Roger Penrose
Collected Works

VOLUME 1

1953–1967

OXFORD SCIENCE PUBLICATIONS

OXFORD
UNIVERSITY PRESS

www.oup.com

ISBN 978 0 19 921936 0

ROGER PENROSE
COLLECTED WORKS
VOLUME 1
1953–1967

常州大学图书馆
藏书章

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York
© Roger Penrose 2011
The moral rights of the authors have been asserted
Crown copyright material is reproduced under Class Licence
Number C01P0000148 with the permission of OPSI
and the Queen's Printer for Scotland
Database right Oxford University Press (maker)
First published 2011
All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above
You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer
British Library Cataloguing in Publication Data
Data available
Library of Congress Control Number: 2010926298
Typeset by Newgen Imaging Systems Pvt Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire
ISBN 978–0–19–921936–0 (Vol. 1)
978–0–19–921944–5 (Set)
1 3 5 7 9 10 8 6 4 2

PREFACE

The topics addressed, in these *Collected Works*, vary over quite an extensive range, the boundaries between their areas being sometimes quite blurred. It would have been hard to classify all these articles and arrange them according to individual topics. Moreover, my own mode of working has been, on the whole, to keep most of my different scientific or mathematical interests running concurrently. Accordingly, it has seemed appropriate not to attempt to group the papers together into separate tidy categories. To do so would have provided a somewhat misleading picture of my overall mode of thinking. Instead, the ordering of articles that has been adopted is essentially chronological, at least with regard to the date of publication. The resulting somewhat higgledy-piggledy appearance that we find in the sometimes abrupt juxtaposition of different topics is, perhaps, fairly illustrative of my actual mode of working.

There are two main kinds of drive that underlie these works. On the one hand, some of the articles have arisen mainly from purely mathematical considerations, which can be ultimately aesthetic or just matters of simple curiosity. On the other hand, motivations for many others of the articles came primarily from a strong desire to discover more about the underlying principles governing the workings of the physical world. Yet, one of the deep mysteries about what has been discovered about these principles is the great breadth of intrinsically appealing mathematics that is found to underlie the actual operation of the physics of the universe, so these two drives are found to relate closely to one another. The reader may perhaps note the distinct geometrical flavour of much of my own mathematics, evidencing the strong appeal that geometry exerts on me, an appeal that is not infrequently intertwined with desire for algebraic elegance. Indeed, I have found it remarkable how much of the workings of our universe appear to depend upon sophisticated and powerful geometry and algebra.

My own very geometrical perspectives, as opposed to purely computational ones, are perhaps made more obvious in the later volumes of this series than in the early ones, but this appearance is deceptive. Very early in my career as a Cambridge graduate student, I had developed a graphical notation which enabled me to carry out many kinds of calculation in an entirely diagrammatic way. For a long time I was rather secretive of this procedure, and in my earlier writings would take care to remove all trace of these visual origins. I do not doubt that my predilections for such geometric or algebraically elegant forms of understanding have influenced the specific choices of direction in which my endeavours towards furthering our understandings of the physics of the world have taken.

The focus of these directions has become more clarified as time has moved on, and has narrowed down to such a degree that there has emerged a somewhat disturbing repetitiveness in many of the later articles. A good number of these provide rather non-technical accounts. An example of such repetitiveness occurs with my descriptions of gravitational collapse to a black hole. There is, however, a gradually developing outlook on this issue, leading to an emphasis on the fundamental contrast in character between the space-time singularities that inevitably appear within black holes and the space-time singularity that constitutes the big bang, this contrast being central to the existence of a second law of thermodynamics in our universe. The same kind of repetitiveness applies to my frequently repeated claim of a need for a fundamental change in the basic laws of quantum theory, where only towards the end do my suggestions become particularly specific. Likewise, many mildly differing accounts are provided outlining the formalism and the basic ideas that underlie twistor theory. The later ones lead up to various viewpoints on the 'googly' problem, which has constituted the main obstruction to the finding of a full understanding of general relativity in twistor terms. Despite such repetition, it is my hope that there is a sufficient variety of emphasis and of historical development in such articles for it to have been worth while to include them all here.

Several colleagues have collaborated with me in writing the works presented here, and I can mention but only a few. Foremost, I have collaborated long and particularly fruitfully with Ezra T. Newman, especially with regard to the asymptotics of general relativity, and also with Wolfgang Rindler in the development of the formalism of 2-spinor calculus. Other collaborators who have played important roles in helping to develop the mathematics of discrete physics and twistor theory have been Erwin Kronheimer, Michael Eastwood, Ronny Wells, and Denny Hill. Among my collaborators, providing very significant inputs, who were former students in my research group, are Andrew Hodges, Paul Tod, Lane Hughston, George Sparling, Richard Ward, Nick Woodhouse, and Lionel Mason, but there were many others. Although not playing roles as actual collaborators, fundamental influences in my research have come from Dennis Sciama and Michael Atiyah.

Roger Penrose
April 2010

BIOGRAPHY

Early years and influences

I was born in Colchester, Essex, UK on 8 August 1931. My father, Lionel Sharples Penrose, FRS, was a highly talented and intelligent man, with many interests. He, himself, had come from a Quaker family, his own father (James Doyle Penrose) having been a professional artist and his mother (Josephine Peckover) was from a well-to-do Quaker banking family. Lionel's primary professional interest was in the causes of hereditary mental disease. He was Galton Professor of Human Genetics at University College London from 1945 and, following retirement, Director of the Kennedy–Galton Institue, St Albans. He had a very considerable artistic talent and a keen interest in music, especially Mozart, but also Bach and other classical composers, dabbling a little in composition himself, particularly when there was a puzzle aspect in what was involved (where he once wrote a piece that could be played equally well upside down as right-way up). He liked making puzzles of various kinds in wood, and spent much of his later life designing wooden models illustrating self-reproduction. He was recognized as a fine composer of chess problems. He had a definite fondness for mathematics and was an innovative statistician. He loved explaining matters of science and mathematics. On one occasion, when I told him that my mathematics teacher at school had informed us that he would be explaining calculus the next day, Lionel immediately took me aside so as to be sure to be the first to teach me about the beauties of calculus. This made a great impression on me. Yet he was an extremely difficult person to communicate with on personal matters.

My mother, born Margaret Leathes, was a charming and caring woman with a very subtle mind. She had been Head Girl at Bedales school, and became medically qualified, but found she was not able to practice after her marriage. After Lionel died, in 1972, she married Max Newman FRS, a fascinating man, who had been Professor of Mathematics in Manchester and, as we later learned, a key figure at Bletchley Park during the war. Margaret's father, John Beresford Leathes, FRS, had been a distinguished professor of physiology, and her mother, a concert pianist, born Sara Mara Natanson from a Jewish family in Latvia, had disowned her family (then in St Petersberg), becoming completely secretive about her origins, and even about her maiden name. Margaret enjoyed geometry at school, but did not

communicate much on this topic with me. My older brother Oliver was very precocious, and I learnt a lot of physics and mathematics from him as I was growing up. He later became a distinguished mathematical physicist (statistical mechanics) and FRS. My younger brother Jonathan excelled at games, particularly chess, and later became British Chess Champion a record 10 times, and then a top correspondence chess player. My much younger sister Shirley Victoria, now Shirley Hodgson, became a highly regarded Professor of Cancer Genetics.

This atmosphere of science, mathematics, art, and music, and of puzzles and games, had been an important part of my upbringing. During the war years, the family lived in London, Ontario, Canada, where Lionel became Director of Psychiatric Research at the Ontario Hospital. We returned to England in 1945, where I attended University College School, London. As an undergraduate, I studied Mathematics at University College London, from 1951 to 1954; then I went to St John's College Cambridge to do research in algebraic geometry, initially under William Hodge and subsequently John Todd. In my same year, under Hodge, was Micheal Atiyah, who was an inspirational, though somewhat intimidating colleague. Frank Adams and Christopher Zeemen (see Chapter 13) were contemporaries who also influenced me, and I became good friends with Hans Liebeck and Douglas Munn (see Chapter 3), who were close research-student colleagues at St John's.

Yet, the colleague who had the greatest influence on the development of my research during this period, and for many years later, was the cosmologist Dennis Sciama. Even before studying at Cambridge, I had a chance meeting with him during a visit to my brother Oliver in Cambridge in which I raised a question that had worried me in relation to the inspiring series of BBC radio talks given by Fred Hoyle in 1951 (The nature of the universe. A series of broadcast lectures). On the strength of this encounter (at which I used a space-time diagram with tilted light cones to make my point), Dennis had evidently formed the opinion that it would be worth while to interest me more in physics, even trying, later, to persuade me to switch my topic of research to cosmology. I was then too committed to my many pure-mathematical activities, but I learnt a great deal of physics from Dennis, and of the excitements involved in striving to uncover secrets of the universe. Another influential colleague was Felix Pirani, from whom I learnt much about developments in general relativity.

Pivotal were the superb lecture courses of Bondi, on general relativity, and Dirac, on quantum mechanics. On the pure-mathematical side, I recall particularly an elegant information-packed algebra course by Philip Hall. Lectures by Steen on mathematical logic turned out to influence me strongly towards my later viewpoint concerning the relation between computation and conscious understanding (e.g. Chapters 169, 184, 186).

Degrees

1952 BSc Mathematics (Spec) (1st Cl), University College, London

1957 PhD (Cantab), St John's College, Cambridge

Employment

1956	NRDC London (research on numerical analysis with computers); later consultant
1956–1957	Assistant Lecturer, Bedford College, London
1957–1960	Research Fellow, St John's College, Cambridge
1959–1961	NATO Research Fellow, Princeton, Syracuse, Cornell Universities
1961–1963	Research Associate, King's College, London
1963–1964	Visiting Associate Professor, University of Texas, Austin, Texas
1964–1966	Reader (Applied Mathematics), Birkbeck College, London
1966–1967	Visiting appointments: Yeshiva, Princeton, Cornell, Chicago Universities; lecturer Battelle Inst., Seattle
1967–1973	Professor (Applied Mathematics), Birkbeck College, London
1969–1970	Visiting professor, Yeshiva New York, Princeton, and Cornell Universities
1973–1998	Rouse Ball Professor (Mathematics), Oxford; 1998– Emeritus
1978	Miller Fellow, Univ. of California, Berkeley
1982–1987	Lovett Professor (Mathematics), Rice Univ., Houston, Texas, USA (Part-time)
1987–1993	Distinguished Professor (Physics, Mathematics), Syracuse Univ. USA (Part-time)
1993–	Francis and Helen Pentz Distinguished Professor (Physics, Mathematics; part-time)
1998–2001	Gresham Professor (Geometry), Gresham College, London

Honours Prizes include 1971 Heinemann Prize; 1972 Fellow of Royal Society; 1978 Plenary Lecture, International Congress of Mathematicians, Helsinki; 1985 Royal Medal (Royal Society);1988 Wolf Prize (physics, with Hawking); 1990 Albert Einstein Medal; 1994 knighted (services to science); 1994 Polish Academy of Sciences; 1998 US National Academy of Sciences; 2000 Order of Merit; 2004 de Morgan Medal (London Mathematical Society); 2001 Royal Irish Academy (Science); 2008 Copley Medal (Royal Society).

Marriages

1958	Joan Isabel Wedge; divorced 1980
1988	Vanessa Dee Thomas

Children

1963	Christopher Shaun
1964	Toby Nicholas
1966	Eric Alexander
2000	Maxwell Sebastian

Honorary degrees

1992	University of New Brunswick, Canada
1993	University of Surrey
1994	University of Bath
1995	University of London
1996	University of Glasgow
1996	University of Essex
1997	University of St Andrews
1998	University of Santiniketan, India
1998	Open University
2002	University of Southampton
2004	University of Waterloo, Canada
2005	University of Leuven, Belgium
2005	Warsaw University
2005	University of Athens
2006	University of York

Books

1972	Techniques of Differential Topology in Relativity (NSF and SIAM)
1984 and 1986	Spinors and Space-Time (Volumes 1 and 2; Cambridge University Press; with Rindler)
1989	The Emperor's New Mind (Oxford University Press 1989)
1994	Shadows of the Mind (Oxford University Press)
1996	The Nature of Space and Time (Princeton University Press; with Hawking)
1997	The Large, the Small and the Human Mind (Cambridge University Press)
2004	The Road to Reality (Jonathan Cape)

INTRODUCTION

On looking through the various articles that have found themselves to be in the period spanned by Volume 1 of this 6-volume set of *Collected Works*, I am struck by the number of them (published between the years 1953 and 1967) that form the basis of most of the research that has occupied me for the rest of my life. The general ideas of the algebra of tensor systems and their diagrammatic representations, as addressed by the two theses (Chapters 6 and 7) and the notation introduced in Chapter 31, have formed an important background to most of my later research. The beauty and power of 2-spinor calculus, which had intrigued me as seeming to lie outside the framework of ordinary tensor algebra, was made clear to me from lectures in Cambridge by Paul Dirac. The magnificence of general relativity was impressed upon me by the lectures of Hermann Bondi, and much of my early published research involved bringing these two spheres of understanding together.

Dennis Sciama's influence had guided me in various ways into many of the enticing mysteries of physics, but I had indeed found myself to be particularly attracted by the geometry of spinors and by the powerful principles underlying Einstein's general relativity. Accordingly, my own distinctive approaches to physical problems have concentrated on 2-spinors, spin-coefficients, and the geometry of light cones (particularly Chapters 11, 12, and 15), and this very much guided me with regard to my own peculiar type of approach to the study of space-time geometry. This, in turn, has taken me in certain other directions, one of which led to singularity theorems (Chapter 20) and to the idea of treating asymptotics in terms of conformal boundaries (Chapter 17, 18, and 22). Intrigued by issues raised by the inter-relations between quantum mechanics and space-time structure I played, early on, with ideas relating to the possibility of a discrete basis to physics (Chapters 28, 31, and 32). But as my understanding of quantum mechanics developed I became increasingly impressed by the power, mystery, and geometric subtlety bestowed by the complex number system, as opposed to that provided by discrete systems, and this led me into twistor theory (Chapter 29)—some of the more significant implications of which have only recently begun to come to light.

ACKNOWLEDGEMENTS

Oxford University Press has attempted to trace the copyright holder of all the papers included in these volumes and apologizes to copyright holders if permission to publish in this form has not been obtained.

The author and publishers thank the following individuals for their permissions:

Acta Physics Polonica B **280**
Adam Hilger Ltd **147**
American Chemical Society **206**
American Institute of Physics
American Journal of Physics **19, 24**
Journal of Maths and Physics **15, 26, 27, 29, 39, 41, 58, 84**
Errata **15**
American Mathematical Society **30, 180**
Bulletin of the American Mathematical Society **144**
Contemporary Mathematics **179**
American Physical Society
Physical Review Letters **16, 20, 23, 302**
Reviews of Modern Physics **21**
Annals of Human Genetics **100**
Asian Journal of Mathematics **281**
Astronomische Gesellschaft **295**

Benjamin **36**
Bibliopolis **175**
Blackwell **80, 166, 186, 215**
Annals of the New York Academy of Sciences **60, 190, 294**
British Broadcasting Company **59**
British Journal of Psychology **9**

Cambridge Entomological Club **243**
Cambridge University Press **6, 7, 46, 104, 138, 157, 172, 173, 176, 195, 218, 220, 227, 230, 240, 290, 303, 304**
Behavioural and Brain Sciences **193, 194, 224**
Proceedings of the Cambridge Philosophical Society **3, 4, 5, 32, 42, 50, 126**
Chapman and Hall/CRC Press **107, 124, 125, 127, 128, 130, 136, 139, 142, 143, 146, 149, 150, 154, 155, 156, 160, 161, 164, 170, 171, 178, 189, 191, 198, 201, 203, 209, 229, 234, 293**
Connected Globe **257**
Consigle Nazionale Della Richerche **145**
Cornell University Press **33**

Éditions du Centre National de la Recherche Scientifique **14**
Elsevier **12, 45, 52, 68, 85, 86, 92, 93, 123, 151, 167, 185, 187, 226, 279**
Escher, M.C. images: The M.C. Escher Company B.V. **218, 249, 306**
Eureka **1, 2**

Folio Society **284**
Freeman **56**

Granta **297**
Gresham College **277**

Halsted Press **148**
Hanser **66, 67**

Imperial College Press **264**
Imprint Academic **235, 245**
Indian Academy of Sciences **244**
Indiana University Press **25**
Institute of Electrical and Electronics Engineers **196**
Institute of Physics **217**
Classical and Quantum Gravity **253, 254, 274, 282**
European Journal of Physics **258**
Journal of Physics A: Mathematical and General **95**
International Astronomical Union
International Astronomical Union Symposium Proceedings Series No. 13, Cosmology **72**

International Atomic Energy Agency
Contemporary Physics **38**
International Centre for Theoretical Physics **53, 68**
International Mathematical Union **286**
International University Bremen **305**
Inter-University Centre for Astronomy and Physics **265**
The Italian Astronomical Society
Memorie della Società Astronomica Italiana **200**

Longman **78, 79, 108, 116, 117, 134, 135, 141, 177, 188**

The Massachusetts Institute of Technology Press **228, 246**

National Institute of Sciences of Ukraine **48**
Nature **8, 44, 49**
Nature (Physical Science) **47**
New Scientist **10, 101**
New York Academy of Sciences **294**
Oxford University Press **55, 69, 121, 122, 165, 178, 184, 197, 202, 208, 216, 237, 270, 272, 276, 291, 299**

Pearson **109**
Penguin **273**
Penrose's Unpublished Papers **31**
Pergamon Press **17**
Pitman **73, 75, 76, 77, 82, 87, 89, 90, 94, 102, 103, 110**
Pontificial University **296**
Princeton University Press **28, 268**
Annals of Math **13, 131**
Proceedings of the International Congress of Mathematicians **115**

Rivista del Nuovo Cimento **40**
Routledge **169**
Royal Astronomical Society **183**
The Royal Society
Proceedings of the Royal Society Series A **22, 35, 97, 105, 106, 132**
Philosophical Transactions of the Royal Society of London. Series A **275, 283**

The Sciences **99**
Scientific American **51**
Simon and Schuster **242**
Springer **61, 63, 65, 83, 91, 98, 111, 133, 168, 204, 207, 231, 255, 263, 278, 292, 298, 300**
Artificial Intelligence Review **205**
Communications in Mathematical Physics **43, 54, 129**
Cosmology and Gravitation **120**
Foundations of Physics **140**
General Relativity and Gravitation **70, 71, 113, 114, 239, 251**
Higher-Order and Symbolic Computation **287**
International Journal of Theoretical Physics **37, 56**
Journal of Statistical Physics **233**
Lecture Notes in Physics **112, 152**
The Mathematical Intelligencer **285**
Taylor and Francis **18**
International Studies in the Philosophy of Science **256**
Surveys in High Energy Physics **119**
Trinity College, Dublin **219**

World Scientific **162, 163, 225, 236, 241, 252, 289, 301**

Universal Academy Press **266**
University of Chicago Press **96, 269**
University of Chittagong **158**
University College London **267**

The University of Oxford (Mathematics Department)
Twistor Newsletter **73, 74, 75, 76, 77, 78, 79, 81, 82, 88, 89, 90, 94, 102, 103, 107, 108, 116, 117, 118, 124, 125, 127, 128, 130, 134, 135, 136, 137, 139, 141, 142, 143, 146, 149, 150, 153, 154, 155, 156, 157, 160, 161, 164, 170, 171, 174, 181, 182, 188, 189, 191, 192, 198, 199, 201, 203, 209, 210, 211, 212, 213, 214, 221, 222, 223, 229, 232, 238, 248, 290, 250, 259, 260, 261, 262, 271, 288**

University of Utah Press **247**

CONTENTS
VOLUME 1

	Introduction	*xxiii*
	1950s	
1	Morley's trisector theorem	1
2	'Construction problem' and 'Word problem'	3
3	A note on inverse semigroups	5
4	A generalized inverse for matrices	11
5	On best approximate solutions of linear matrix equations	21
6	Tensor methods in algebraic geometry	25
7	Geometrical algebras: A new approach to invariant theory	251
8	A self-reproducing analogue	351
9	Impossible objects: A special type of visual illusion	355
10	Puzzles for Christmas	359
11	The apparent shape of a relativistically moving sphere	365
	1960s	
12	A spinor approach to general relativity	369
13	Imbedding of manifolds in Euclidean space	403
14	General relativity in spinor form	415
15	An approach to gravitational radiation by a method of spin coefficients	421
16	Asymptotic properties of fields and space-times	437
17	The light cone at infinity	441
18	Conformal treatment of infinity	447
19	Energy conservation as the basis of relativistic mechanics	469
20	Gravitational collapse and space-time singularities	475
21	A remarkable property of plane waves in general relativity	479
22	Zero rest-mass fields including gravitation: Asymptotic behaviour	487
23	10 exact gravitationally conserved quantities	533
24	Energy conservation as the basis of relativistic mechanics II	537
25	General-relativistic energy flux and elementary optics	541
26	Note on the Bondi-Metzner-Sachs group	559

27 Some new gravitationally conserved quantities 569
28 An analysis of the structure of space-time 579
29 Twistor algebra 731
30 Conserved quantities and conformal structure in general relativity 755
31 Theory of quantized directions 769
32 On the structure of causal spaces 801

DETAILED CONTENTS
VOLUMES 1–6

VOLUME 1

1950s

1 Morley's trisector theorem 1
2 'Construction problem' and 'Word problem' 3
3 A note on inverse semigroups 5
4 A generalized inverse for matrices 11
5 On best approximate solutions of linear matrix equations 21
6 Tensor methods in algebraic geometry 25
7 Geometrical algebras: A new approach to invariant theory 251
8 A self-reproducing analogue 351
9 Impossible objects: A special type of visual illusion 355
10 Puzzles for Christmas 359
11 The apparent shape of a relativistically moving sphere 365

1960s

12 A spinor approach to general relativity 369
13 Imbedding of manifolds in Euclidean space 403
14 General relativity in spinor form 415
15 An approach to gravitational radiation by a method of spin coefficients 421
16 Asymptotic properties of fields and space-times 437
17 The light cone at infinity 441
18 Conformal treatment of infinity 447
19 Energy conservation as the basis of relativistic mechanics 469
20 Gravitational collapse and space-time singularities 475
21 A remarkable property of plane waves in general relativity 479
22 Zero rest-mass fields including gravitation: Asymptotic behaviour 487
23 10 exact gravitationally conserved quantities 533
24 Energy conservation as the basis of relativistic mechanics II 537
25 General-relativistic energy flux and elementary optics 541
26 Note on the Bondi-Metzner-Sachs group 559
27 Some new gravitationally conserved quantities 569
28 An analysis of the structure of space-time 579
29 Twistor algebra 731
30 Conserved quantities and conformal structure in general relativity 755
31 Theory of quantized directions 769
32 On the structure of causal spaces 801

VOLUME 2

1960s

33 Cosmological boundary conditions for zero rest-mass fields 1
34 On gravitational collapse and cosmology 15
35 New conservation laws for zero rest-mass fields in asymptotically flat space-time 29
36 Structure of space-time 61
37 Twistor quantization and curved space-time 177
38 On gravitational collapse 217

39 Solutions of the zero-rest-mass equations 231
40 Gravitational collapse: The role of general relativity 235
41 Conserved quantities in the Einstein–Maxwell theory 261

1970s

42 The singularities of gravitational collapse and cosmology 267
43 On quadratic first integrals of the geodesic equations for type {22} spacetimes 289
44 Scattering of two impulsive gravitational plane waves 301
45 Applications of negative dimensional tensors 307
46 Angular momentum: An approach to combinatorial space-time 333
47 Extraction of rotational energy from a black hole 367
48 Spinor classification of energy tensors 375
49 Black holes and gravitational theory 389
50 Ideal points in space-time 395
51 Black holes 419
52 Twistor theory: An approach to the quantisation of fields and space-time 429
53 a) 'Radiation, energy and the BMS group'; b) 'Some recent developments'; c) 'N-P constants' 505
54 On a quadratic first integral for the charged particle orbits in the charged Kerr solution 513
55 The geometry of impulsive gravitational waves 521
56 On the nature of quantum geometry 537
57 Internal instability in a Reissner–Nordström black hole 583
58 A space-time calculus based on pairs of null directions 599
59 Black holes 609
60 Naked singularities 635
61 Relativistic symmetry groups 647
62 Gravitational collapse 707
63 Gravitational collapse 729
64 The role of aesthetics in pure and applied mathematical research 741
65 Singularities in cosmology 749
66 Twistors and particles: An outline 761
67 The non-linear graviton 779
68 Classical gravity: A status report 791
69 Twistor theory: Its aims and achievements 795

VOLUME 3

1970s

70 Nonlinear gravitons and curved twistor theory 1
71 The nonlinear graviton 25
72 The space-time singularities of cosmology and in black holes 33
73 The twistor quadrille 45
74 The snail contour 53
75 The universal bracket factor 55
76 Some elementary twistor integrals with boundary 59
77 Twistor functions and sheaf cohomology 63
78 How to grow (or at least half grow) new twistors from old 77
79 Some remarks on the two twistor description of electromagnetic and linearized gravitational potentials 81
80 The cosmic censorship hypothesis 83
81 A finite covering for PT^+ 87
82 The back-handed photon (or, to a cricketer, the googly photon!) 89
83 Any space-time has a plane wave as a limit 99
84 The Kähler structure of asymptotic twistor space 105
85 The twistor programme 113
86 Space-time singularities 127
87 The twisted camel (or how to get through the eye of a singularity). Includes: more on the twisted camel 137
88 Remarks concerning the ambidextrous photon 145
89 Massless fields and sheaf cohomology 149
90 Note on the n-twistor internal symmetry group 159
91 Aspects of general relativity 163
92 Is nature complex? 203
93 Gravitational collapse: A review 213
94 Local $H^{1,}$s and propagation 231

95 Extended Regge trajectories 237
96 Singularities of spacetime 243
97 The metric and curvature properties of $\mathcal{H}$-space 271
98 The geometry of the universe 297
99 Einstein's vision and the mathematics of the natural world 341
100 The topology of ridge systems 347
101 Twisting round space-time 359
102 A googly graviton? 365
103 Twistors as helicity raising operators 375
104 Singularities and time-asymmetry 381
105 General introduction 441
106 Global general relativity 445
107 Remarks on the Sparling and Eguchi-Hanson (googly?) gravitons 449
108 The anti-self-dual Coulomb field's non-Hausdorff twistor space 459
109 On the twistor descriptions of massless fields 461
110 Combinatorial quantum theory and quantized directions 499
111 Pentaplexity: A class of nonperiodic tilings of the plane 507

1980s

112 Recent advances in global general relativity: A brief survey 515
113 Twistors for flat and curved space-time 527
114 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity 575
115 The complex geometry of the natural world 617
116 On the evaluation of twistor cohomology classes 625
117 Three channels for the box diagram 629
118 Puzzle: A sequence for twistorians 637
119 A brief introduction to twistors 639
120 A brief outline of twistor theory 663

VOLUME 4

1980s

121 Time-asymmetry and quantum gravity 1
122 Some remarks on twistors and curved-space quantization 31
123 On Schwarzschild causality—a problem for "Lorentz covariant" general relativity 47
124 Concerning a Fourier contour integral 61
125 A new angle on the googly graviton 65
126 A quadratic mapping with invariant cubic curve 73
127 Twistors for cosmological models 91
128 Physical left-right symmetry and googlies 97
129 Cohomology and massless fields 105
130 A theory of 2-surface ('superficial') twistors 153
131 Some unsolved problems in classical general relativity 161
132 Quasi-local mass and angular momentum in general relativity 201
133 Time-asymmetry, cosmological uniformity and space-time singularities 213
134 Real-space topology for advanced/retarded twistor functions 217
135 Twistor blisters and relative cohomology 221
136 On the geometry of googly maps 225
137 A puzzle for twistorians 229
138 Some remarks on gravity and quantum mechanics 231
139 General-relativistic kinematics? 241
140 Spinors and torsion in general relativity 247
141 Non-locality in physics: A new mathematical viewpoint 263
142 On closed-set coverings for PT^+—A correction 269
143 Remarks on curved-space, twistor theory and googlies 271
144 A note on Sparling's 3-form 277
145 Physical space-time and nonrealizable CR-structures 281
146 Quasi-local mass and angular momentum 305
147 Relative cohomology, googlies and deformations of $\mathbb{I}$ 309
148 Donaldson's moduli space: A 'model' for quantum gravity? 313
149 Mass in general relativity 319
150 'New improved' quasi-local mass, and the Schwarzschild solution 331
151 On Bryant's condition for holomorphic curves in CR space 339

152 Integrals for general-relativistic sources: A development from Maxwell's electromagnetic theory 343
153 Mass and angular momentum at the quasi-local level in general relativity 377
154 Objective state-vector reduction? 387
155 Embedding 2-surfaces in CM 393
156 A suggested further modification to the quasi-local formula 397
157 Higher-dimensional two-surface twistors 399
158 Numerical relativity by power series 403
159 Elements of twistor theory 409
160 Energy and its definition in general relativity 429
161 Two-surface twistors and hypersurface twistors 441
162 Exponentiating a relative H^2 445
163 Gravitational field mass 449
164 Gravity and quantum theory 467
165 Twistors and the time-irreversability of state-vector reduction 473
166 Gravity and state vector reduction 477
167 Big bangs, black holes and 'time's arrow' 497
168 Escher and the visual representation of mathematical ideas 525
169 Hermann Weyl, space-time and conformal geometry 541
170 Quantum physics and conscious thought 573
171 Local twistor transport at $\mathscr{I}^+$: An approach to the googly 601
172 An approach to a coordinate-free calculus at $\mathbb{I}$ 609
173 Twistors in general relativity 613
174 Newton, quantum theory and reality 633
175 On Qadir's intertwining operators 667
176 On the origins of twistor theory 671
177 New directions in quantum gravity 693
178 Pretzel twistor spaces 703
179 Twistors and state-vector reduction 713
180 Aspects of quasi-local angular momentum 717
181 Fundamental asymmetry in physical laws 727
182 Topological QFT and twistors: Holomorphic linking 741
183 Holomorphic linking: Postscript 745
184 Singularities and big-bang cosmology 747
185 On the physics and mathematics of thought 751
186 A twistor conformal field theory for four space-time dimensions 785
187 Minds, machines and mathematics 791
188 Tilings and quasi-crystals: A non-local growth problem? 811
189 Non-Hausdorff Riemann surfaces and complex dynamical systems 839
190 Using spinors in general relativity 845
191 A twistorial approach to the full vacuum equations 851

VOLUME 5

1990s

192 Difficulties with inflationary cosmology 1
193 Light rays near i^0: A new mass-positivity theorem 19
194 Extended Regge trajectories: 12 years after 27
195 Précis of *The Emperor's New Mind: Concerning computers, minds and the laws of physics* 29
196 Author's response – the nonalgorithmic mind 43
197 Twistor theory after 25 years – its physical status and prospects 59
198 Twistors, particles, strings and links 89
199 Mass positivity from focussing and the structure of space-like infinity 101
200 Twistor theory for vacuum space-time: A new approach 107
201 What does the big bang tell us about quantum gravity? 113
202 Twistors as spin 3/2 charges 123
203 Twistors as charges for spin 3/2 in vacuum 133
204 The mass of the classical vacuum 143
205 Twistors as spin 3/2 charges continued: SL(3,C) bundles 151
206 Response to Tony Dodd's 'Godel, Penrose and the possibility of AI' 159
207 The biggest enigma 161
208 $\mathcal{H}$-space and twistors 167
209 'Penrose and mathematical ability' 189

210 The most general {2,2} self-dual vacuum: A googly approach 193
211 On impulsive gravitational waves 199
212 Kinking and causality 205
213 Puzzle page 213
214 Solution to puzzle in TN34 215
215 Gravitational collapse of the wave function: New thoughts 221
216 Setting the scene: The claim and the issues 225
217 The modern physicist's view of nature 259
218 Gravity and quantum mechanics 311
219 Foreword to *Mr Tompkins in paperback* 323
220 Mathematics, physics, computers and conscious understanding 327
221 An example of indeterminacy in the time-development of "already unified field theory": A collision between electromagnetic plane waves 339
222 Orthogonality of general spin states 349
223 A simple observation concerning {2 2} vacuums 355
224 The orthographic image of a regular tetrahedron 357
225 Author's response – An Emperor still without mind 361
226 Square-root voting 369
227 On Bell non-locality without probabilities: More curious geometry 375
228 Quantum non-locality and complex reality 401
229 On the cohomology of impossible figures 415
230 Spin 3/2 fields and local twistors 419
231 Mathematical intelligence 427
232 Is conscious awareness consistent with space-time descriptions? 459
233 A twistor-topological approach to the Einstein equations 475
234 On the second law of thermodynamics 485
235 Twistors and the Einstein equations 491
236 What 'gaps'? Reply to Grush and Churchland 505
237 Non-locality and objectivity in quantum state reduction 521
238 Must mathematical physics be reductionist? 531
239 Concerning space-time points for spin 3/2 twistor spaces 547
240 In {2 2} vacuums, the null-datum on null rays is exactly r^{-5} 553
241 Why new physics is needed to understand the mind 557
242 The role of gravity in quantum state reduction 573
243 Consciousness involves noncomputable ingredients 593
244 Beyond the doubting of a shadow: A reply to commentaries on *Shadows of the Mind* 603
245 Chandrasekhar, black holes, and singularities 641
246 Conscious events as orchestrated space-time selections 661
247 Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness? 681
248 Space-time and cosmology 711
249 New ideas for the googly graviton 757
250 Incidence between complex null rays 771
251 Two puzzles 777
252 On gravity's role in quantum state reduction 779
253 Twistor theory, the Einstein equations, and quantum mechanics 801

VOLUME 6

1990s

254 Andrzej Mariusz Trautman 1
255 Twistor geometry of light rays 11
256 Remarks on tiling: Details of a $(1+\varepsilon+\varepsilon^2)$-aperiodic set 37
257 On understanding understanding 69
258 Mathematics: Pattern, precision and profundity 85
259 The mathematics of the electron's spin 109
260 Five things you can do with a surface in PT 115
261 Solutions to puzzles in TN41 121
262 Corrections to missprints in solutions to TN41 puzzles 125
263 Googly maps as forms 127
264 The need for a non-computational extension of quantum action in the brain 133
265 Chandrasekhar and the end of time 151
266 Science and the mind 161

267 Quantum state-reduction as a gravitational phenomenon; theory and experiment 165
268 Lionel Penrose: Colleague and father 169
269 Foreword to *Einstein's miraculous year: Five papers that changed the face of physics* 175
270 The question of cosmic censorship 185
271 Afterword to *The geometric universe: Science, geometry, and the work of Roger Penrose* 203
272 The twistor space for a strongly asymptotically flat vacuum 213
273 The complexity of our singular universe 223
274 Can a computer understand? 251
275 Spherically-symmetric solutions of the Schrödinger–Newton equations 279
276 Quantum computation, entanglement and state reduction 291
277 Foreword to *Ludwig Boltzmann: The man who trusted atoms* 305
278 Mathematics in the arts 309
279 Some remarks on twistor theory 323
280 The central programme of twistor theory 339
281 Lecture in honour of Leopold Infeld (extended outline only): Spinors in general relativity 371
282 On Ricci-flat twistor theory 381
283 Twistor theory and the Einstein vacuum 389
284 Introduction: Science at the turn of the millennium. 'Science into the next millennium: Young scientists give their visions of the future, part I.' 409

2000s

285 Introduction. *What is life?* 417
286 Drawing with complex numbers 433
287 Mathematical physics of the 20th and 21st centuries 441
288 Reminiscences of Christopher Strachey 459
289 On extracting the googly information 463
290 Wavefunction collapse as a real gravitational effect 489
291 On Bell non-locality without probabilities: Some curious geometry 509
292 The heritage of Pythagoras: Nineteen to the dozen 537
293 Twistors and general relativity 551
294 Towards a twistor description of general space-times: Introductory comments 579
295 Consciousness, the brain, and spacetime geometry: An addendum 597
296 The Schwarzschild singularity 605
297 The physics of the small and large: What is the bridge between them? 621
298 The rediscovery of gravity. The Einstein equation of general relativity 629
299 John Bell, state reduction, and quanglement 663
300 Consciousness, computation, and the Chinese room 677
301 Twistor geometry of conformal infinity 703
302 Gravitational collapse of the wavefunction: An experimentally testable proposal 713
303 Towards quantum superpositions of a mirror 719
304 On the instability of extra space dimensions 725
305 The problem of spacetime singularities: Implications for quantum gravity? 743
306 The remarkable role of mathematics in the physical universe 767

1
MORLEY'S TRISECTOR THEOREM

Eureka **16** (1953), *6–7*

Morley's Trisector Theorem

A WELL-KNOWN theorem due to F. Morley states that the three points in which pairs of internal trisectors of a triangle nearest a given side meet, are the vertices of an equilateral triangle.

There are many proofs of this theorem. None of these, however, appears to be quite as simple as the following, which is a good example of the value of "working backwards from the answer." p.[6]

We shall prove that given any triangle with angles 3α, 3β, 3γ we can construct the required configuration *starting with the equilateral triangle*. (Clearly it is only the angles and not the size that are important.)

Let PQR be an equilateral triangle. Denote the angle PQR by pQr, etc. Choose L outside the triangle PQR such that

$$lQr = lRq = \beta + \gamma$$

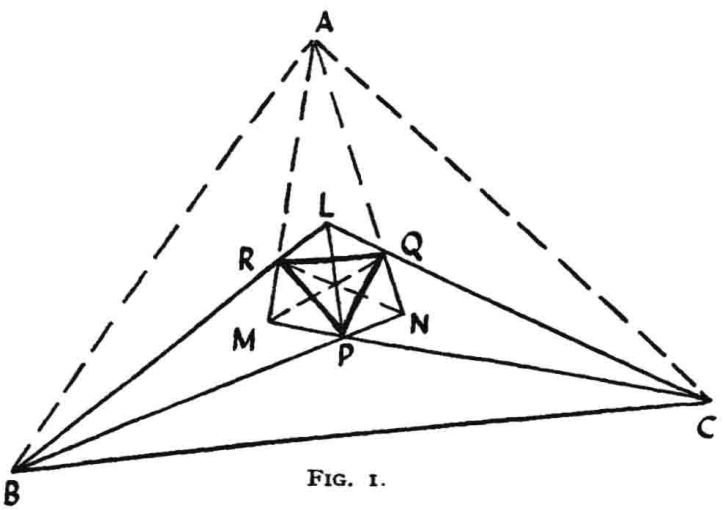


FIG. 1.

Clearly LP bisects qLr. M and N are similarly defined. Let LQ meet MP in C. Then

$$qCp = lQp + mPq - \pi = \alpha + \beta + 2\gamma - \pi/3 = \gamma$$

since $$\alpha + \beta + \gamma = \pi/3$$

Choose B on LR so that CP bisects qCb p.[7]

P is the in-centre of triangle LCB and so

$$pBr = pBc = \tfrac{1}{2}(\pi - rLq - lCb) = \beta$$

giving $$bPr + rPn = \pi$$

Hence B is the intersection of LR and NP. Similarly if A is the intersection of MR and NQ we have

$$qCa = \gamma;\ qAr = qAc = \alpha$$

also $$rBa = \beta;\ rAb = \alpha$$

and so ABC is the required triangle and the theorem is proved.

R. P.

2
'CONSTRUCTION PROBLEM' AND 'WORD PROBLEM'

Eureka **17** (1954), 13

The construction problem posed here arose out of a study that I had made whilst an undergraduate (in the period 1949–50) of a kind of synthetic geometry in which the basic elements could be represented as points and unit circles in a Euclidean plane (with a duality between them). About a year later I found a highly symmetrical geometrical theorem involving eight conics in a plane, from which most of the well-known configurations of plane projective geometry (including those of Pascal, Brianchon, and Poncelet) could be derived through specialization. I intended to publish these, but have not, even now, got around to it.

Word Problem. (*Solution on page* 17.)

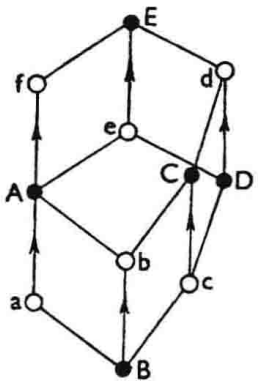
From the four relations ABA = A, BAB = B, AB = BA, AC = CA prove formally that BC = CB. (R. PENROSE.)

Construction Problem. (*Solution on page* 17.)

Using only a circular coin of unit radius, one is allowed to construct circles of unit radius through any two given points which are not more than a distance of 2 apart. The problem is to construct a circle to touch another such circle at a given point. (R. PENROSE.)

Solution of Construction Problem. (*Page* 13.)

In the diagram, the black spots represent intersections of circles and the white ones the centres of circles. The lines joining the spots are of unit length. If *a* and A are respectively the given circle and the given point, we can construct the figure in the following manner. We choose B arbitrarily on *a* and construct *b*, then C on *b*



to give *c*, and D (not too far from A) on *c* to give *d* and *e*. This determines E and we can construct *f*. The lines with arrows are parallel.

Solution to Word Problem. (*Page* 13.)

Written in full, this requires 17 steps. We have

$$BC=B^2AC=B^2CA=B^2CA^3B^2=B^2A^3CB^2=ACB^2=CAB^2=CB$$

3
A NOTE ON INVERSE SEMIGROUPS

Proc. Camb. Phil. Soc. **51** (1955), 396–399. With W. D. Munn

Reprinted from the *Proceedings of the Cambridge Philosophical Society*,
Volume 51, Part 2, pp. 396-399, 1955.
PRINTED IN GREAT BRITAIN

A NOTE ON INVERSE SEMIGROUPS

By W. D. MUNN and R. PENROSE

Received 13 November 1954

The concept of an 'inverse semigroup' or 'generalized group' has recently been introduced, independently, by Preston (4) and Vagner (5). In the present note we give several alternative definitions for such a system, some of which appear to be weaker than that originally given.

A *semigroup* is a non-empty set S which is closed with respect to an associative binary operation. An element e of S is termed an *idempotent* if $e^2 = e$. A *left ideal* L of S is a non-empty subset of S such that $SL \subseteq L$, and a *right ideal* R of S is a non-empty subset of S such that $RS \subseteq R$. A *two-sided ideal* (or, simply, an *ideal*) of S is a subset which is both a left and a right ideal.

Following Green (3), we shall say that the elements a, b of S are $\mathfrak{l}$-*equivalent* or $\mathfrak{r}$-*equivalent* according as a and b generate the same left or right ideal* of S respectively, and we write $a \equiv b(\mathfrak{l})$ or $a \equiv b(\mathfrak{r})$. If a left or right ideal is generated by a single element, it is termed *principal*.

A semigroup S is said to be *regular* if for any given element $a \in S$ there is at least one solution of the equation $axa = a$ in S. We observe that in a regular semigroup the equations

$$axa = a, \quad xax = x$$

have at least one common solution. For suppose u is any solution of $axa = a$. Write $v = uau$. Then $ava = auaua = aua = a$, and $vav = uauauau = uau = v$; hence v is a common solution of the given equations.

An *inverse semigroup* is a regular semigroup in which all the idempotents commute. This is essentially the definition given by Preston and Vagner. Alternative definitions are provided by the following result:

THEOREM. *The following conditions on a semigroup S are equivalent:*

C1. *S is regular, $\mathfrak{l}$-equivalent idempotents commute, and $\mathfrak{r}$-equivalent idempotents commute.*

C2. *Each principal left ideal and each principal right ideal of S is generated by one and only one idempotent.*†

C3. *S is regular, and for any idempotent $e \in S$ the equations $exe = e$, $xex = x$ have no common idempotent solution other than e.*

C4. *For any element $a \in S$ the equations $axa = a$, $xax = x$ have a unique common solution.*

C5. *S is an inverse semigroup.*

* For example, the left ideal of S generated by a is $Sa \cup a$.

† That the left and right properties of S in C1 and C2 are in fact independent can be seen by considering a semigroup S whose multiplication is given by $ab = b$ for all $a, b \in S$.

We require five lemmas.

LEMMA 1. *If two* l*-equivalent* (r*-equivalent*) *idempotents of a semigroup S commute, then they are equal.*

Proof. Let e, f be l-equivalent idempotents. Then there exists $u \in S$ such that $e = uf$. Hence $ef = uf^2 = uf = e$. Similarly $fe = f$. Thus if $ef = fe$ we have $e = f$. The proof for r-equivalent idempotents is similar.

LEMMA 2. *If each principal left* (*right*) *ideal of a semigroup S is generated by at least one idempotent, then S is regular. Conversely each principal left ideal and each principal right ideal of a regular semigroup is generated by an idempotent.*

(This is Theorem 6 of (3), but we give the proof for the sake of completeness.)

Proof. Suppose that for a given $a \in S$ we can find an idempotent e such that $e \equiv a(\mathfrak{l})$. Then there exist elements $u, v \in S$ such that $e = ua$, $a = ve$. Hence $ae = ve^2 = ve = a$ and so $aua = a$. A similar argument holds for principal right ideals.

Conversely if u is any solution of the equation $axa = a$, then $ua \equiv a(\mathfrak{l})$ and $au \equiv a(\mathfrak{r})$. But ua and au are idempotents and the result follows.

LEMMA 3. *If each principal left ideal and each principal right ideal of a semigroup S is generated by one and only one idempotent, then for any* $a \in S$ *the equations* $axa = a$, $xax = x$ *have a unique common solution.*

Proof. By Lemma 2, S is regular. Let u, v be common solutions of $axa = a$, $xax = x$. Then we have $ua \equiv a(\mathfrak{l})$, $va \equiv a(\mathfrak{l})$ and so $ua \equiv va(\mathfrak{l})$. But ua and va are idempotents and so $ua = va$. Similarly $au = av$. Thus we have

$$u = (ua)u = (va)u = v(au) = v(av) = v,$$

as required.

LEMMA 4. *Let S be a regular semigroup in which for any given idempotent e the equations* $exe = e$, $xex = x$ *have no common idempotent solution other than e. Then each principal left ideal and each principal right ideal is generated by one and only one idempotent.*

Proof. Since S is regular it follows from Lemma 2 that each principal left ideal and each principal right ideal is generated by at least one idempotent. Let e and f be l-equivalent idempotents. Then, as in the proof of Lemma 1, we have $ef = e$, $fe = f$. Hence $efe = e^2 = e$ and $fef = f^2 = f$. Thus from our hypothesis we have $e = f$. Similarly, if e and f are r-equivalent idempotents, then $e = f$.

LEMMA 5. *Let S be a semigroup in which for any given element a the equations* $axa = a$, $xax = x$ *have a unique common solution. Then S is an inverse semigroup.*

Proof. Let e and f be any two idempotents of S, and let g be the unique common solution of the equations

$$(ef)x(ef) = ef, \quad x(ef)x = x. \tag{1}$$

Then $$(ef)fg(ef) = (ef)g(ef) = ef \quad \text{and} \quad fg(ef)fg = f.g(ef)g = fg,$$

whence, by the uniqueness of g, we have $g = fg$. Similarly, it can be shown that $g = ge$. Thus $g^2 = (ge)(fg) = g$, and so g is a solution of the equations $xgx = x$, $gxg = g$. But ef is also a solution of these equations since g is the solution of equations (1).

Hence $g = ef$, which shows that ef is an idempotent. In the same way we can show that fe is an idempotent.

Thus $(ef)fe(ef) = (ef)^2 = ef$ and $fe(ef)fe = (fe)^2 = fe$,

and so fe is a solution of equations (1). Hence $ef = fe$ as required.

We now complete the proof of the theorem. Lemmas 1 and 2 show that C 1 implies C 2. Also C 2 implies C 4 by Lemma 3, and C 4 implies C 5 by Lemma 5. But *a fortiori* C 5 implies C 1, and so C 1, C 2, C 4, C 5 are equivalent. Further C 3 implies C 2 by Lemma 4, and *a fortiori* C 4 implies C 3. Hence C 1, C 2, C 3, C 4, C 5 are equivalent.

Preston and Vagner have proved that C 5 implies C 4.

The proof of Lemma 5 is a generalization of part of that given by Croisot to establish the result that a semigroup in which the equation $axa = a$ has a unique solution for any a, must be a group ((2), Theorem 5). We give a stronger result:

COROLLARY 1. *A regular semigroup in which for any idempotent e the equation $exe = e$ has no idempotent solution other than e is a group.*

Proof. Let S be a semigroup with the above property. Then the equations $exe = e$, $xex = x$ have the unique common idempotent solution e for any given idempotent e of S; that is, S satisfies C 3. Hence S is an inverse semigroup. Thus if e and f are any two idempotents of S, then ef is an idempotent, and so the equation $(ef)x(ef) = ef$ has a unique idempotent solution. But e and f are such solutions, and so $e = f$. Finally, let a be any element of S, and let u be any solution of $axa = a$. Then $au = ua = e$, the unique idempotent of S, and $ea = ae = a$. Hence S is a group.

A *semilattice* is a commutative semigroup of idempotents. Now let S be a semigroup which is expressible as a union of disjoint groups G_α, where α belongs to a semilattice Y. Then Clifford (1) has termed S a *semilattice of groups* if the multiplication in S is such that $G_\alpha G_\beta \subseteq G_{\alpha\beta}$, where $\alpha\beta$ is the product of the elements α, β in Y. We then have

COROLLARY 2. *If S is a semilattice of groups, then all the idempotents of S commute.*

Proof. With the above notation we see that two elements are l-equivalent (r-equivalent) if and only if they belong to the same group G_α, say. But G_α contains one and only one idempotent, and so S satisfies C 2. Thus S is an inverse semigroup.

This result was previously obtained by Clifford ((1), Theorem 8) and by Croisot ((2), p. 375).

Let M be an ideal of a semigroup S. Then the semigroup $S - M$ is defined to be the set obtained from S by collapsing M to a single element 0, say, the remaining elements of S being unchanged. Multiplication ($\circ$) in $S - M$ is then defined by the rules

(i) $a \circ b = ab$ if $a, b \neq 0$ and $ab \notin M$,

(ii) $a \circ b = 0$ if $a, b \neq 0$ and $ab \in M$,

and (iii) $a \circ 0 = 0 = 0 \circ a$ for all $a \in S - M$.

COROLLARY 3. *Let M be an ideal of a semigroup S. Then S is an inverse semigroup if and only if M and $S - M$ are inverse semigroups.*

Proof. We use C 4. Let S be an inverse semigroup and a any element of S. Let u be the unique common solution of the equations $axa = a$, $xax = x$. Then $u \in M$ if and only if $a \in M$. It follows that M and $S - M$ also satisfy C 4.

Conversely suppose M and $S-M$ satisfy C4. Then if a is any element of S, the equations $axa = a$, $xax = x$ have a common solution u, say. But $u \in M$ if and only if $a \in M$. Hence u is uniquely determined by a; that is, S satisfies C4.

REFERENCES

(1) Clifford, A. H. Bands of semigroups. *Proc. Amer. math. Soc.* 5 (1954), 499–504.
(2) Croisot, R. Demi-groupes inversifs et demi-groupes réunions de demi-groupes simples. *Ann. sci. Éc. norm. sup., Paris*, (3), 70 (1953), 361–79.
(3) Green, J. A. On the structure of semigroups. *Ann. Math., Princeton* (2), 54 (1951), 163–72.
(4) Preston, G. B. Inverse semi-groups. *J. Lond. math. Soc.* 29 (1954), 396–403.
(5) Vagner, V. V. Generalized groups. *C.R. Acad. Sci. U.R.S.S.* 84 (1952), 1119–22 (Russian).

St John's College
Cambridge

Note added in Proof. The result of Lemma 5 has also recently been obtained by A. E. Liber, On the theory of generalized groups. *C.R. Acad. Sci. U.R.S.S.* 97 (1954), 26–28 (Russian).

4
A GENERALIZED INVERSE FOR MATRICES

Proc. Camb. Phil. Soc. **51** (1955), 406–413

I found this notion of 'generalized inverse' (now frequently referred to as a 'pseudoinverse') as a consequence of trying to solve the main problem that had been the focus of my Ph.D. thesis (Chapter 6). This had led me to the matrix equation AXA = A (actually in a group ring, not directly matrices) and I asked my older brother Oliver if he knew under what circumstances it could be solved. His response, 'maybe it always has a solution', was the starting point of this paper. I realized afterwards that I should have acknowledged him in the paper—perhaps this mention of his help now is some sort of recompense! I was not aware of E.H. Moore's very much earlier discovery of this same concept (done somewhat differently) until this was pointed out to me by Richard Rado (see Chapter 5). I denoted the generalized inverse of A by $A^{\dagger}$. For some reason the notation A^{+} tends to be used nowadays, which seems to me to be most inappropriate. This concept is now generously referred to as the Moore–Penrose pseudoinverse. I intended a follow-up paper on various algebraic aspects of generalized inverses, but never got round to it.

Reprinted from the *Proceedings of the Cambridge Philosophical Society*,
Volume 51, Part 3, pp. 406–413, 1955.
PRINTED IN GREAT BRITAIN

A GENERALIZED INVERSE FOR MATRICES

By R. PENROSE

Communicated by J. A. Todd

Received 26 July 1954

This paper describes a generalization of the inverse of a non-singular matrix, as the unique solution of a certain set of equations. This generalized inverse exists for any (possibly rectangular) matrix whatsoever with complex elements‡. It is used here for solving linear matrix equations, and among other applications for finding an expression for the principal idempotent elements of a matrix. Also a new type of spectral decomposition is given.

In another paper its application to substitutional equations and the value of hermitian idempotents will be discussed.

Notation. Capital letters always denote matrices (not necessarily square) with complex elements. The conjugate transpose of the matrix A is written A^*. Small letters are used for column vectors (with an asterisk for row vectors) and small Greek letters for complex numbers.

The following properties of the conjugate transpose will be used:

$$A^{**} = A,$$

$$(A+B)^* = A^* + B^*,$$

$$(\lambda A)^* = \bar{\lambda} A^*,$$

$$(BA)^* = A^* B^*,$$

$$AA^* = 0 \textit{ implies } A = 0.$$

The last of these follows from the fact that the trace of AA^* is the sum of the squares of the moduli of the elements of A. From the last two we obtain the rule

$$BAA^* = CAA^* \textit{ implies } BA = CA, \tag{1}$$

since $$(BAA^* - CAA^*)(B-C)^* = (BA-CA)(BA-CA)^*.$$

Similarly $$BA^*A = CA^*A \textit{ implies } BA^* = CA^*. \tag{2}$$

Theorem 1. *The four equations*

$$AXA = A, \tag{3}$$

$$XAX = X, \tag{4}$$

$$(AX)^* = AX, \tag{5}$$

$$(XA)^* = XA, \tag{6}$$

have a unique solution for any A.

‡ Matrices over more general rings will be considered in a later paper.

Proof. I first show that equations (4) and (5) are equivalent to the single equation

$$XX^*A^* = X. \tag{7}$$

Equation (7) follows from (4) and (5), since it is merely (5) substituted in (4). Conversely, (7) implies $AXX^*A^* = AX$, the left-hand side of which is hermitian. Thus (5) follows, and substituting (5) in (7) we get (4). Similarly, (3) and (6) can be replaced by the equation

$$XAA^* = A^*. \tag{8}$$

Thus it is sufficient to find an X satisfying (7) and (8). Such an X will exist if a B can be found satisfying

$$BA^*AA^* = A^*.$$

For then $X = BA^*$ satisfies (8). Also, we have seen that (8) implies $A^*X^*A^* = A^*$ and therefore $BA^*X^*A^* = BA^*$. Thus X also satisfies (7).

Now the expressions A^*A, $(A^*A)^2$, $(A^*A)^3, \ldots$ cannot all be linearly independent, i.e. there exists a relation

$$\lambda_1 A^*A + \lambda_2(A^*A)^2 + \ldots + \lambda_k(A^*A)^k = 0, \tag{9}$$

where $\lambda_1, \ldots, \lambda_k$ are not all zero. Let λ_r be the first non-zero λ and put

$$B = -\lambda_r^{-1}\{\lambda_{r+1}I + \lambda_{r+2}A^*A + \ldots + \lambda_k(A^*A)^{k-r-1}\}.$$

Thus (9) gives $B(A^*A)^{r+1} = (A^*A)^r$, and applying (1) and (2) repeatedly we obtain $BA^*AA^* = A^*$, as required.

To show that X is unique, we suppose that X satisfies (7) and (8) and that Y satisfies $Y = A^*Y^*Y$ and $A^* = A^*AY$. These last relations are obtained by respectively substituting (6) in (4) and (5) in (3). (They are (7) and (8) with Y in place of X and the reverse order of multiplication and must, by symmetry, also be equivalent to (3), (4), (5) and (6).) Now

$$X = XX^*A^* = XX^*A^*AY = XAY = XAA^*Y^*Y = A^*Y^*Y = Y.$$

The unique solution of (3), (4), (5) and (6) will be called the generalized inverse of A (abbreviated g.i.) and written $X = A^\dagger$. (Note that A need not be a square matrix and may even be zero.) I shall also use the notation $\lambda^\dagger$ for scalars, where $\lambda^\dagger$ means λ^{-1} if $\lambda \neq 0$ and 0 if $\lambda = 0$.

In the calculation of $A^\dagger$ it is only necessary to solve the two unilateral linear equations $XAA^* = A^*$ and $A^*AY = A^*$. By putting $A^\dagger = XAY$ and using the fact that XA and AY are hermitian and satisfy $AXA = A = AYA$ we observe that the four relations $AA^\dagger A = A$, $A^\dagger AA^\dagger = A^\dagger$, $(AA^\dagger)^* = AA^\dagger$ and $(A^\dagger A)^* = A^\dagger A$ are satisfied. Relations satisfied by $A^\dagger$ include

$$\left.\begin{aligned} A^\dagger A^{\dagger *}A^* = A^\dagger = A^*A^{\dagger *}A^\dagger \\ \text{and} \qquad A^\dagger AA^* = A^* = A^*AA^\dagger, \end{aligned}\right\} \tag{10}$$

these being (7), (8) and their reverses.

Note that the properties $(A+B)^* = A^*+B^*$ and $(\lambda A)^* = \bar{\lambda}A^*$ were not used in the proof of Theorem 1.

LEMMA. 1·1 $A^{\dagger\dagger} = A$;

1·2 $A^{*\dagger} = A^{\dagger *}$;

1·3 *if A is non-singular* $A^\dagger = A^{-1}$;

1·4 $(\lambda A)^\dagger = \lambda^\dagger A^\dagger$;

1·5 $(A^*A)^\dagger = A^\dagger A^{\dagger *}$;

1·6 *if U and V are unitary* $(UAV)^\dagger = V^*A^\dagger U^*$;

1·7 *if* $A = \Sigma A_i$, *where* $A_i A_j^* = 0$ *and* $A_i^* A_j = 0$, *whenever* $i \neq j$ *then* $A^\dagger = \Sigma A_i^\dagger$;‡

1·8 *if A is normal* $A^\dagger A = AA^\dagger$ *and* $(A^n)^\dagger = (A^\dagger)^n$;

1·9 A, A^*A, $A^\dagger$ *and* $A^\dagger A$ *all have rank equal to* trace $A^\dagger A$.

Proof. Since the g.i. of a matrix is always unique, relations 1·1, ..., 1·7 can be verified by substituting the right-hand side of each in the defining relations for the required g.i. in each case. For 1·5 we require (10) and for 1·7 we require the fact that $A_i A_j^\dagger = 0$ and $A_i^\dagger A_j = 0$ if $i \neq j$. This follows from $A_j^\dagger = A_j^* A_j^{\dagger *} A_j^\dagger$ and $A_i^\dagger = A_i^\dagger A_i^{\dagger *} A_i^*$. The first part of 1·8, of which the second part is a consequence, follows from 1·5 and (10), since $A^\dagger A = (A^*A)^\dagger A^*A$ and $AA^\dagger = (AA^*)^\dagger AA^*$. (The condition for A to be normal is $A^*A = AA^*$.)‡ To prove 1·9 we observe that each of A, A^*A, $A^\dagger$ and $A^\dagger A$ can be obtained from the others by pre- and post-multiplication by suitable matrices. Thus their ranks must all be equal and $A^\dagger A$ being idempotent has rank equal to its trace.

Relation 1·5 provides a method of calculating the g.i. of a matrix A from the g.i. of A^*A, since $A^\dagger = (A^*A)^\dagger A^*$. Now A^*A, being hermitian, can be reduced to diagonal form by a unitary transformation. Hence $A^*A = UDU^*$, where U is unitary and $D = \text{diag}\,(\alpha_1, \ldots, \alpha_n)$. Thus $D^\dagger = \text{diag}\,(\alpha_1^\dagger, \ldots, \alpha_n^\dagger)$ and $(A^*A)^\dagger = UD^\dagger U^*$ by 1·6.

An alternative proof of the existence of the g.i. of a square matrix can be obtained from Autonne's (1) result that any square matrix A can be written in the form VBW, where V and W are unitary and B is diagonal§. Now $B^\dagger$ exists since it can be defined similarly to $D^\dagger$ above. It is thus sufficient to verify that $A^\dagger = W^*B^\dagger V^*$ as in 1·6. A rectangular matrix can be treated by bordering it with zeros to make it square.

Notice that $A^\dagger$ is a continuous function of A if the rank of A is kept fixed, since in the singular case the polynomial in (9) could be taken to be the characteristic function of A^*A. The value of r, being the nullity of A^*A (as A^*A is hermitian), is thus determined by the rank of A (by 1·9). Also $A^\dagger$ varies discontinuously when the rank of A changes since rank A = trace $A^\dagger A$ by 1·9.

‡ This result, like several of the earlier ones, is a simple consequence of the construction of $A^\dagger$ given in Theorem 1. It seems desirable, however, to show that it is also a direct consequence of (3), (4), (5) and (6).

§ A simple proof of this result is as follows: since AA^* and A^*A are both hermitian and have the same eigenvalues there exists a unitary matrix T such that $TAA^*T^* = A^*A$. Hence TA is normal and therefore diagonable by unitary transformation.

THEOREM 2. *A necessary and sufficient condition for the equation* $AXB = C$ *to have a solution is*

$$AA^{\dagger}CB^{\dagger}B = C,$$

in which case the general solution is

$$X = A^{\dagger}CB^{\dagger} + Y - A^{\dagger}AYBB^{\dagger},$$

where Y *is arbitrary.*

Proof. Suppose X satisfies $AXB = C$. Then

$$C = AXB = AA^{\dagger}AXBB^{\dagger}B = AA^{\dagger}CB^{\dagger}B.$$

Conversely, if $C = AA^{\dagger}CB^{\dagger}B$ then $A^{\dagger}CB^{\dagger}$ is a particular solution of $AXB = C$ (i.e. we may treat $A^{\dagger}$ formally as an inverse of A in order to find a particular solution).

For the general solution we must solve $AXB = 0$. Now any expression of the form $X = Y - A^{\dagger}AYBB^{\dagger}$ satisfies $AXB = 0$ and conversely if $AXB = 0$ then $X = X - A^{\dagger}AXBB^{\dagger}$. Theorem 2 follows. Note that the only property required of $A^{\dagger}$ for Theorem 2 is $AA^{\dagger}A = A$.

COROLLARY 1. *The general solution of the vector equation* $Px = c$ *is*

$$x = P^{\dagger}c + (I - P^{\dagger}P)y,$$

where y *is arbitrary, provided that the equation has a solution.*‡

The general linear matrix equation $A_{(1)}XB_{(1)} + \ldots + A_{(m)}XB_{(m)} = C$ can also be solved by using a g.i., since this equation can be written in the form

$$\sum_{j,k}\left\{\sum_{r=1}^{m}\alpha_{(r)ij}\beta_{(r)kl}\right\}\xi_{jk} = \gamma_{il},$$

where $A_{(r)} = (\alpha_{(r)ij})$, etc. If X and C are regarded as vectors, this is an equation of the type considered in Corollary 1.

COROLLARY 2. *A necessary and sufficient condition for the equations* $AX = C$, $XB = D$ *to have a common solution is that each equation should individually have a solution and that* $AD = CB$.

Proof. The condition is obviously necessary. To show that it is sufficient, put

$$X = A^{\dagger}C + DB^{\dagger} - A^{\dagger}ADB^{\dagger},$$

which is a solution if the required conditions $AA^{\dagger}C = C$, $DB^{\dagger}B = D$, $AD = CB$ are satisfied.

This result is due to Cecioni (3).

Hermitian idempotents. A *hermitian idempotent* matrix (abbreviated h.i.) is one satisfying $EE^{*} = E$, that is, $E = E^{*}$ and $E^{2} = E$.

LEMMA. 2·1 $A^{\dagger}A$, $AA^{\dagger}$, $I - A^{\dagger}A$, $I - AA^{\dagger}$ *are all h.i.*;

2·2 *if* E *is h.i.* $E^{\dagger} = E$;

2·3 K *is idempotent if and only if there exist h.i.'s* E *and* F *such that* $K = (FE)^{\dagger}$ *in which case* $K = EKF$.

‡ See also Bjerhammar (2).

Proof. The proofs of 2·1 and 2·2 are evident. To prove 2·3 suppose $K^2 = K$, then, by 1·1, $K = \{(K^\dagger K)(KK^\dagger)\}^\dagger$. Thus $E = KK^\dagger$ and $F = K^\dagger K$ will do. Conversely, if $K = (FE)^\dagger$ then K is of the form $EFPEF$ (since $Q^\dagger = Q^*(Q^{\dagger *}Q^\dagger Q^{\dagger *})Q^*$ by (10)). Hence $K = EKF$, so that $K^2 = E(FE)^\dagger FE(FE)^\dagger F = E(FE)^\dagger F = K$.

Principal idempotent elements of a matrix. For any square matrix A there exists a unique set of matrices K_λ defined for each complex number λ such that

$$K_\lambda K_\mu = \delta_{\lambda\mu} K_\lambda, \tag{11}$$

$$\Sigma K_\lambda = I, \tag{12}$$

$$AK_\lambda = K_\lambda A, \tag{13}$$

$$(A - \lambda I)K_\lambda \text{ is nilpotent}, \tag{14}$$

the non-zero K_λ's being the *principal idempotent elements* of A (see Wedderburn (7)).‡ Unless λ is an eigenvalue of A, K_λ is zero so that the sum in (12) is finite, though summed over all complex λ. The following theorem gives an 'explicit' formula for K_λ in terms of g.i.'s.

THEOREM 3. *If*

$$E_\lambda = I - \{(A-\lambda I)^n\}^\dagger (A-\lambda I)^n \quad \text{and} \quad F_\lambda = I - (A-\lambda I)^n \{(A-\lambda I)^n\}^\dagger,$$

where n is sufficiently large (e.g. the order of A), then the principal idempotent elements of A are given by $K_\lambda = (F_\lambda E_\lambda)^\dagger$, and n can be taken as unity if and only if A is diagonable.

Proof. First, suppose A is diagonable and put

$$E_\lambda = I - (A-\lambda I)^\dagger (A-\lambda I)$$

and

$$F_\lambda = I - (A-\lambda I)(A-\lambda I)^\dagger.$$

Then E_λ and F_λ are h.i. by 2·1 and are zero unless λ is an eigenvalue of A by 1·3.

Now

$$(A-\mu I)E_\mu = 0 \quad \text{and} \quad F_\lambda(A-\lambda I) = 0, \tag{15}$$

so that $\mu F_\lambda E_\mu = F_\lambda A E_\mu = \lambda F_\lambda E_\mu$. Thus

$$F_\lambda E_\mu = 0 \quad \text{if} \quad \lambda \neq \mu. \tag{16}$$

Putting $K_\lambda = (F_\lambda E_\lambda)^\dagger$ we have, by 2·3,

$$K_\lambda = E_\lambda (F_\lambda E_\lambda)^\dagger F_\lambda. \tag{17}$$

Hence $K_\lambda K_\mu = \delta_{\lambda\mu} K_\lambda$ by (16). Also (17) gives

$$F_\lambda K_\mu E_\nu = \delta_{\lambda\mu}\delta_{\mu\nu} F_\lambda E_\lambda. \tag{18}$$

Any eigenvector z_α of A corresponding to the eigenvalue α satisfies $E_\alpha z_\alpha = z_\alpha$. Since A is diagonable, any column vector x conformable with A is expressible as a sum of eigenvectors; i.e. it is expressible in the form (a finite sum over all complex λ)

$$x = \Sigma E_\lambda x_\lambda.$$

‡ The existence of such K_λ's can be established as follows: let $\phi(\xi) \equiv (\xi-\lambda_1)^{n_1} \dots (\xi-\lambda_r)^{n_r}$ be, say, the minimum function of A where the factors $(\xi-\lambda_i)^{n_i}$ are mutually coprime. Then if $\phi(\xi) \equiv (\xi-\lambda_i)^{n_i}\psi_i(\xi)$ there exist polynomials $\chi_i(\xi)$ such that $\Sigma\chi_i(\xi)\psi_i(\xi) \equiv 1$. Set $K_{\lambda_i} = \chi_i(A)\psi_i(A)$ with the other K_λ's zero.

Similarly, if y^* is conformable with A, it is expressible as

$$y^* = \Sigma y_\lambda^* F_\lambda.$$

Now
$$\begin{aligned} y^*(\Sigma K_\mu)x &= (\Sigma y_\lambda^* F_\lambda)(\Sigma K_\mu)(\Sigma E_\nu x_\nu) \\ &= \Sigma y_\lambda^* F_\lambda E_\lambda x_\lambda = (\Sigma y_\lambda^* F_\lambda)(\Sigma E_\nu x_\nu) \\ &= y^* x \end{aligned}$$

by (18) and (16). Hence $\Sigma K_\mu = I$. Also, from (15) and (17) we obtain

$$AK_\lambda = \lambda K_\lambda = K_\lambda A. \tag{19}$$

Thus conditions (13) and (14) are satisfied. Also

$$A = \Sigma \lambda K_\lambda. \tag{20}$$

Conversely, if it is not assumed that A is diagonable but that $\Sigma K_\lambda = I$, K_λ being defined as above (i.e. with $n = 1$), then $x = \Sigma K_\lambda x$ gives x as a sum of eigenvectors of A, since (19) was deduced without assuming the diagonability of A.

If A is not diagonable it seems more convenient simply to prove that for any set of K_λ's satisfying (11), (12), (13) and (14) each $K_\lambda = (F_\lambda E_\lambda)^\dagger$, where F_λ and E_λ are defined as in the theorem.

Now, if the K_λ's satisfy (11), (12), (13) and (14) they must certainly satisfy

$$\Sigma K_\lambda = I$$

and
$$(A - \lambda I)^n K_\lambda = 0 = K_\lambda (A - \lambda I)^n,$$

where n is sufficiently large (if n is the order of A, this will do).

From this and the definitions of E_λ and F_λ given in the theorem we obtain

$$E_\lambda K_\lambda = K_\lambda = K_\lambda F_\lambda. \tag{21}$$

Using Euclid's algorithm we can find P and Q (polynomials in A) such that

$$I = (A - \lambda I)^n P + Q(A - \mu I)^n,$$

provided that $\lambda \neq \mu$. Now $F_\lambda (A - \lambda I)^n = 0 = (A - \mu I)^n E_\mu$. Hence $F_\lambda E_\mu = 0$ if $\lambda \neq \mu$, which with (21) gives $F_\lambda K_\mu = 0 = K_\lambda E_\mu$ whenever $\lambda \neq \mu$. But $\Sigma K_\lambda = I$. Thus

$$\left.\begin{aligned} F_\lambda K_\lambda &= F_\lambda \\ \text{and} \quad K_\lambda E_\lambda &= E_\lambda. \end{aligned}\right\} \tag{22}$$

Using (21) and (22) it is a simple matter to verify that $(F_\lambda E_\lambda)^\dagger = K_\lambda$. (Note that $K_\lambda = I - \{(I - E_\lambda)(I - F_\lambda)\}^\dagger$ also.)

Incidentally, this provides a proof of the uniqueness of K_λ.‡

COROLLARY. *If A is normal, it is diagonable and its principal idempotent elements are hermitian.*

Proof. Since $A - \lambda I$ is normal we can apply 1·8 to the definitions of E_λ and F_λ in Theorem 3. This gives

$$E_\lambda = I - (A - \lambda I)^\dagger (A - \lambda I) \quad \text{and} \quad F_\lambda = I - (A - \lambda I)(A - \lambda I)^\dagger,$$

whence A is diagonable. Also $E_\lambda = F_\lambda$, so that $K_\lambda = E_\lambda$ is hermitian.

‡ In fact, $\Sigma K_\lambda = I$ and $(A - \lambda I)^n K_\lambda = 0$ are sufficient for uniqueness.

For this result see, for example, Drazin (4).

If A is normal (20) gives $A = \Sigma\lambda E_\lambda$, and thus using 1·7, 1·4 and 2·2 we obtain $A^\dagger = \Sigma\lambda^\dagger E_\lambda$.

A new type of spectral decomposition. Unless A is normal, there does not appear to be any simple expression for $A^\dagger$ in terms of its principal idempotent elements. However, this difficulty is overcome with the following decomposition, which applies equally to rectangular matrices.

THEOREM 4. *Any matrix A is uniquely expressible in the form*

$$A = \sum_{\alpha>0} \alpha U_\alpha,$$

this being a finite sum over real values of α, where $U_\alpha^\dagger = U_\alpha^*$, $U_\alpha^* U_\beta = 0$ and $U_\alpha U_\beta^* = 0$ if $\alpha \neq \beta$.‡

Proof. The condition on the U's can be written comprehensively as

$$U_\alpha U_\beta^* U_\gamma = \delta_{\alpha\beta}\delta_{\beta\gamma} U_\alpha.$$

For $U_\alpha U_\alpha^* U_\alpha = U_\alpha$ implies $U_\alpha^* U_\alpha U_\alpha^* = U_\alpha^*$, whence by uniqueness, $U_\alpha^\dagger = U_\alpha^*$ (as $U_\alpha U_\alpha^*$ and $U_\alpha^* U_\alpha$ are both hermitian). Also $U_\alpha U_\alpha^* U_\beta = 0$ and $U_\alpha U_\beta^* U_\beta = 0$ respectively imply $U_\alpha^* U_\beta = 0$ and $U_\alpha U_\beta^* = 0$ by (2). Define

$$E_\lambda = I - (A^*A - \lambda I)^\dagger (A^*A - \lambda I).$$

The matrix A^*A is normal, being hermitian, and is non-negative definite. Hence the non-zero E_λ's are its principal idempotent elements and $E_\lambda = 0$ unless $\lambda \geqslant 0$.

Thus $$A^*A = \Sigma\lambda E_\lambda \quad \text{and} \quad (A^*A)^\dagger = \Sigma\lambda^\dagger E_\lambda.$$

Hence $$A^\dagger A = (A^*A)^\dagger A^*A = \Sigma\lambda^\dagger\lambda E_\lambda = \sum_{\lambda>0} E_\lambda.$$

Put $U_\alpha = \alpha^{-1} A E_{\alpha^2}$ if $\alpha > 0$ and $U_\alpha = 0$ otherwise. Then

$$\Sigma\alpha U_\alpha = \sum_{\alpha>0} \alpha\alpha^{-1} A E_{\alpha^2} = A \sum_{\lambda>0} E_\lambda = AA^\dagger A = A.$$

Also if $\alpha, \beta, \gamma > 0$

$$\begin{aligned} U_\alpha U_\beta^* U_\gamma &= \alpha^{-1}\beta^{-1}\gamma^{-1} A E_{\alpha^2} E_{\beta^2} A^* A E_{\gamma^2} \\ &= \alpha^{-1}\beta^{-1}\gamma^{-1}\gamma^2 A E_{\alpha^2} E_{\beta^2} E_{\gamma^2} \\ &= \delta_{\alpha\beta}\delta_{\beta\gamma}\alpha^{-1} A E_{\alpha^2} = \delta_{\alpha\beta}\delta_{\beta\gamma} U_\alpha. \end{aligned}$$

To prove the uniqueness suppose

$$A = \sum_{\alpha>0} \alpha V_\alpha, \quad \text{where} \quad V_\alpha V_\beta^* V_\gamma = \delta_{\alpha\beta}\delta_{\beta\gamma} V_\alpha.$$

Then it is easily verified that the non-zero expressions $V_\alpha^* V_\alpha$, for $\alpha > 0$, and $I - \sum_{\beta>0} V_\beta^* V_\beta$ are the principal idempotent elements of A^*A corresponding respectively to the eigenvalues α^2 and 0. Hence $V_\alpha^* V_\alpha = E_{\alpha^2}$, where $\alpha > 0$, giving

$$U_\alpha = \alpha^{-1}\Big(\sum_{\beta>0} \beta V_\beta\Big) V_\alpha^* V_\alpha = V_\alpha.$$

Polar representation of a matrix. It is a well-known result (closely related to Autonne's) that any square matrix is the product of a hermitian with a unitary matrix (see, for example, Halmos (5)).

‡ Thus $A^\dagger = \Sigma\alpha^\dagger U_\alpha^*$.

Put $H = \Sigma\alpha U_\alpha U_\alpha^*$. Then H is non-negative definite hermitian, since $U_\alpha = 0$ unless $\alpha > 0$, and $H^2 = AA^*$. (Such an H must be unique.) We have $H^\dagger H = HH^\dagger = AA^\dagger$.

Now $AA^\dagger$ and $A^\dagger A$ are equivalent under a unitary transformation, both being hermitian and having the same eigenvalues; i.e. there is a unitary matrix W satisfying $WA^\dagger A = AA^\dagger W$. Putting $V = H^\dagger A + W - WA^\dagger A$ we see that $VV^* = I$ and $A = HV$. This is the *polar representation* of A.

The polar representation is unique if A is non-singular but not if A is singular. However, if we require $A = HU$, where H is (as before) non-negative definite hermitian, $U^\dagger = U^*$ and $UU^* = H^\dagger H$, the representation is always unique‡ and also exists for rectangular matrices. The uniqueness of H follows from

$$AA^* = HUU^*H = H(H^\dagger H)H = H^2,$$

and that of U from $$H^\dagger A = H^\dagger HU = UU^\dagger U = U.$$

The existence of H and U is established by $H = \Sigma\alpha U_\alpha U_\alpha^*$ and $U = \Sigma U_\alpha$.

If we put $G = \Sigma\alpha U_\alpha^* U_\alpha$ we obtain the alternative representation $A = UG$.

I wish to thank Dr M. P. Drazin for useful suggestions and for advice in the preparation of this paper.

REFERENCES

(1) Autonne, L. Sur les matrices hypohermitiennes et sur les matrices unitaires. *Ann. Univ. Lyon*, (2), 38 (1915), 1–77.
(2) Bjerhammar, A. Rectangular reciprocal matrices, with special reference to geodetic calculations. *Bull. géod. int.* (1951), pp. 188–220.
(3) Cecioni, F. Sopra operazioni algebriche. *Ann. Scu. norm. sup. Pisa*, 11 (1910), 17–20.
(4) Drazin, M. P. On diagonable and normal matrices. *Quart. J. Math.* (2), 2 (1951), 189–98.
(5) Halmos, P. R. *Finite dimensional vector spaces* (Princeton, 1942).
(6) Murray, F. J. and von Neumann, J. On rings of operators. *Ann. Math., Princeton*, (2), 37 (1936), 141–3.
(7) Wedderburn, J. H. M. *Lectures on matrices* (Colloq. Publ. Amer. math. Soc. no. 17, 1934).

St John's College
Cambridge

‡ See also Murray and von Neumann (6). In fact $U^\dagger = U^*$ follows from $UU^* = H^\dagger H$.

5
ON BEST APPROXIMATE SOLUTIONS OF LINEAR MATRIX EQUATIONS

Proc. Camb. Phil. Soc. **52** (1956), 17–19

Reprinted from the *Proceedings of the Cambridge Philosophical Society*,
Volume 52, Part 1, pp. 17–19, 1956.
PRINTED IN GREAT BRITAIN

ON BEST APPROXIMATE SOLUTIONS OF LINEAR MATRIX EQUATIONS

By R. PENROSE

Communicated by J. A. Todd

Received 15 February 1955

In an earlier paper (4)‡ it was shown how to define for any matrix a unique generalization of the inverse of a non-singular matrix. The purpose of the present note is to give a further application which has relevance to the statistical problem of finding 'best' approximate solutions of inconsistent systems of equations by the method of least squares. Some suggestions for computing this generalized inverse are also given.

Notation. In this paper capital letters always denote matrices (square or rectangular) with real or complex elements, the conjugate transpose of the matrix A (or simply the transpose if A is real) being written A^*. The shapes of matrices occurring in expressions are assumed unrestricted except by considerations of conformability. The symbol I denotes any suitable unit matrix.

The generalized inverse of A is defined (as in (4)) to be the unique matrix $A^\dagger$ satisfying

$$AA^\dagger A = A, \quad A^\dagger AA^\dagger = A^\dagger, \quad (AA^\dagger)^* = AA^\dagger, \quad (A^\dagger A)^* = A^\dagger A.$$

If A is real, so also is $A^\dagger$, since it is unique; if A is non-singular, then $A^\dagger = A^{-1}$.

I shall make use of the following relations, all of which are to be found in (4), but which are also immediate consequences of the above:

$$A^*AA^\dagger = A^*, \tag{1}$$

$$A^{\dagger\dagger} = A, \tag{2}$$

$$A^\dagger AA^\dagger = A^\dagger, \tag{3}$$

$$A^*A^{\dagger *}A^\dagger = A^\dagger. \tag{4}$$

I use the notation $\| A \|$ for the sum of the squares of the moduli of the elements of A. Clearly $\| A \| = \operatorname{trace} A^*A$, and $\| A \| > 0$ unless $A = 0$.

Definition. I shall say that X_0 is a *best approximate solution* of the equation $f(X) = G$ if for all X, either

$$\text{(i)} \quad \| f(X) - G \| > \| f(X_0) - G \|,$$

or

$$\text{(ii)} \quad \| f(X) - G \| = \| f(X_0) - G \| \quad \text{and} \quad \| X \| \geqslant \| X_0 \|.$$

THEOREM. *$A^\dagger B$ is the unique best approximate solution of the equation $AX = B$.*

Proof. We have the identity

$$\| AP + (I - AA^\dagger)Q \| = \| AP \| + \| (I - AA^\dagger)Q \|, \tag{5}$$

‡ See also Moore (3) and a forthcoming note by R. Rado.

since (1) implies

$$\{AP+(I-AA^\dagger)Q\}^*\{AP+(I-AA^\dagger)Q\} = (AP)^*AP+\{(I-AA^\dagger)Q\}^*(I-AA^\dagger)Q.$$

Thus, in particular,

$$\begin{aligned}\|AX-B\| &= \|A(X-A^\dagger B)+(I-AA^\dagger)(-B)\| \\ &= \|AX-AA^\dagger B\|+\|AA^\dagger B-B\| \geqslant \|A(A^\dagger B)-B\|,\end{aligned}$$

with equality only when $AX = AA^\dagger B$. Replacing A by $A^\dagger$ in (5) and using (2), we deduce $\|A^\dagger B+(I-A^\dagger A)X\| = \|A^\dagger B\|+\|(I-A^\dagger A)X\|$. Thus, if $AX = AA^\dagger B$, (3) gives $\|X\| = \|A^\dagger B\|+\|X-A^\dagger B\|$, from which the theorem follows.

COROLLARY 1. *The general linear matrix equation* $A_{(1)}XC_{(1)}+\ldots+A_{(m)}XC_{(m)} = B$ *has a unique best approximate solution.*

Proof. This equation is only apparently more general than that of the theorem, since it can be written as

$$\sum_{\beta,\gamma}\left(\sum_{r=1}^{m} a_{(r)\alpha\beta}c_{(r)\gamma\delta}\right)x_{\beta\gamma} = b_{\alpha\delta},$$

where $A_{(r)} = (a_{(r)\alpha\beta})$, etc. This is of the form $RX = S$, where X and S are vectors.

In particular, the equation $AXC = B$ can be treated in this way, the best approximate solution being $A^\dagger BC^\dagger$. This result can also be obtained from the identity

$$\|APC+Q-AA^\dagger QC^\dagger C\| = \|APC\|+\|Q-AA^\dagger QC^\dagger C\|.$$

COROLLARY 2. *The best approximate solution of* $AX = I$ *is* $A^\dagger$.

This shows that $A^\dagger$ is the only matrix having all those properties required of it for the theorem and suggests a possible alternative definition of $A^\dagger$.

The equation $AX = B$, where X and B are vectors, is of special interest in statistics, its best approximate solution being that required in the method of least squares when the equation has no actual solution (see (5)). It would appear, therefore, that a suitable method of computing $A^\dagger$ from the matrix A might be of value. In most practical cases, however, the matrix A^*A is non-singular, so that $A^\dagger = (A^*A)^{-1}A^*$. This gives nothing different from the conventional method of solution (see also (1)).

It is possible to calculate $A^\dagger$ even when A^*A and AA^* are both singular by either of the following methods.

Method 1. Any matrix can be partitioned in the form

$$\begin{pmatrix} A & B \\ C & CA^{-1}B \end{pmatrix}$$

(using a suitable arrangement of rows and columns), A being any non-singular submatrix whose rank is equal to that of the whole matrix. It is then easily verified that

$$\begin{pmatrix} A & B \\ C & CA^{-1}B \end{pmatrix}^\dagger = \begin{pmatrix} A^*PA^* & A^*PC^* \\ B^*PA^* & B^*PC^* \end{pmatrix},$$

where $P = (AA^*+BB^*)^{-1}A(A^*A+C^*C)^{-1}$. The matrices AA^*+BB^* and A^*A+C^*C are positive definite, since A is non-singular. Thus the generalized inverse of any matrix can be expressed in terms of ordinary reciprocals of matrices.

Method 2. The generalized inverse of a given matrix A can be obtained from any matrix D satisfying $A^*A = D(A^*A)^2$, since multiplication on the right by $A^\dagger A^{\dagger *} A^\dagger$ gives $A^\dagger = DA^*$ (from (1) and (4)). Now put $B = A^*A = (b_{\alpha\beta})$, and define a sequence of matrices $C_{(j)}$ for $j = 1, 2, \ldots$ by

$$\left.\begin{aligned} C_{(1)} &= I, \\ C_{(j+1)} &= I.\frac{1}{j}\operatorname{trace}(C_{(j)}B) - C_{(j)}B \quad (j = 1, 2, \ldots). \end{aligned}\right\} \tag{6}$$

I shall show that $C_{(r+1)}B = 0$ and $\operatorname{trace}(C_{(r)}B) \neq 0$, where r is the rank of B, so that if $D = \dfrac{rC_{(r)}}{\operatorname{trace}(C_{(r)}B)}$, then $DB^2 = B$ as required.

Let $c_{(1)\alpha\beta} = \delta_{\alpha\beta}$, $c_{(j+1)\alpha\beta} = \dfrac{1}{j!}\sum_{\gamma_1,\ldots,\gamma_j} \begin{vmatrix} b_{\gamma_1\gamma_1} & \cdots & b_{\gamma_1\gamma_j} & \delta_{\gamma_1\beta} \\ \vdots & & \vdots & \vdots \\ b_{\gamma_j\gamma_1} & \cdots & b_{\gamma_j\gamma_j} & \delta_{\gamma_j\beta} \\ b_{\alpha\gamma_1} & \cdots & b_{\alpha\gamma_j} & \delta_{\alpha\beta} \end{vmatrix}$ $(j = 1, 2, \ldots)$.

The summation extends over all the ordered sets $\gamma_1, \ldots, \gamma_j$; so that each non-zero term of the sum is repeated $j!$ times. Expanding by Laplace's rule with respect to the last column, we get

$$c_{(j+1)\alpha\beta} = \delta_{\alpha\beta}\frac{1}{j!}\sum_{\gamma_1,\ldots,\gamma_j}\begin{vmatrix} b_{\gamma_1\gamma_1} & \cdots & b_{\gamma_1\gamma_j} \\ \vdots & & \vdots \\ b_{\gamma_j\gamma_1} & \cdots & b_{\gamma_j\gamma_j} \end{vmatrix} - j\frac{1}{j!}\sum_{\gamma_1,\ldots,\gamma_{j-1}}\begin{vmatrix} b_{\gamma_1\gamma_1} & \cdots & b_{\gamma_1\gamma_{j-1}} & b_{\gamma_1\beta} \\ \vdots & & \vdots & \vdots \\ b_{\gamma_{j-1}\gamma_1} & \cdots & b_{\gamma_{j-1}\gamma_{j-1}} & b_{\gamma_{j-1}\beta} \\ b_{\alpha\gamma_1} & \cdots & b_{\alpha\gamma_{j-1}} & b_{\alpha\beta} \end{vmatrix}.$$

Comparing this with (6) we see that $C_{(j)} = (c_{(j)\alpha\beta})$. Now if B is of rank r, $C_{(r+1)}B$ is zero, since its elements are sums of minors of B of order $r+1$. Also, $\operatorname{trace} C_{(r)}B$ is not zero, since it is r times the sum of the principal minors of B of order r, and B is a positive semi-definite matrix.

Thus, relation (6) affords a repetitive method of computing $A^\dagger$, while the definition of $c_{(j)}$ leads to an explicit form for $A^\dagger$ when the rank of A^*A (i.e. of A) is known.

If the matrix B is replaced by any non-singular matrix, the above argument still holds, giving $D = B^{-1}$. This is essentially Frame's method of computing the inverse of a matrix (see (2)).

REFERENCES

(1) Bjerhammar, A. Rectangular reciprocal matrices with special reference to geodetic calculations. *Bull. géod. int.* (1951), pp. 188–220.
(2) Dwyer, P. S. *Linear computations* (New York, 1951), pp. 225–8.
(3) Moore, E. H. *Bull. Amer. math. Soc.* (2) 26 (1920), 394–5.
(4) Penrose, R. A generalized inverse for matrices. *Proc. Camb. phil. Soc.* 51 (1955), 406–13.
(5) Turnbull, H. W. and Aitken, A. C. *Theory of canonical matrices* (London, 1948), pp. 173–4.

St John's College
Cambridge

6
TENSOR METHODS IN ALGEBRAIC GEOMETRY

PhD Thesis, Cambridge 1957

Hodge suggested the Ph.D. problem of finding (in the general case) the Cayley form of the intersection of two (unmixed) algebraic varieties, whose Cayley forms (now commonly called 'Chow forms') are given. Initially, it seemed there should be a polynomial solution, and I realized that this could be understood as a tensor manipulation problem. Developing a diagrammatic notation to make such manipulations accessible to me in a visual way, I realized that such a solution generally did not exist. A solution in terms of rational functions looked very plausible, however, but I was never able to prove its correctness generally. These problems took me in various different directions, resulting in this very unwieldy thesis. One such direction led me to define a general algebraic theory of abstract tensor systems, but unfortunately I made a bad (though then natural-looking) choice with regard to notation, making the thesis virtually unreadable. (The approaches of Chapter 31 and 36 eventually rectified this problem.) Parts A and B of the thesis were intended to be supplemented by a Part C, which I never got around to writing, describing the general sets of points that can arise from finite logical operations applied to finite collections of algebraic varieties.

TENSOR METHODS IN ALGEBRAIC GEOMETRY

By

ROGER PENROSE

A dissertation in part fulfilment of the
requirements for Ph.D. degree.

St John's College,
Cambridge.

September, 1956

SUMMARY

A new technique is described for treating certain problems connected with algebraic geometry and invariant theory. The technique developed involves a systematic use of tensor algebra. The use of coordinates is generally avoided and, instead, each geometrical object is associated with a tensor, the geometrical properties of these objects being interpreted in the formal algebraic rules for manipulating tensors. A tensor technique such as this is in a sense equivalent to the classical methods of invariant theory. However the symbolic calculus of the classical theory is not well suited to the treatment of the non-general forms considered here.

The work is in two parts. Section A is concerned with algebraic properties of tensors treated from a purely abstract point of view. Various algebraic restrictions are placed on the tensor systems and one of these is shown to lead to a system isomorphic with that of the "ordinary" tensors. Linear tensor equations are treated, considerable use being made of the "generalized inverse" of a tensor (analogous to the "general reciprocal" of E. H. Moore).

In section B, these tensor methods are applied to the geometry of a projective space. Among the applications given is a tensor proof of the first fundamental theorem of invariant theory. It is shown how this can

CHAPTER 6

be made to lead at once, via a theorem due to Hodge on complexes of linear spaces, to the basis theorem for polynomials in Grassmann coordinates and the second fundamental theorem of invariant theory. A technique for treating non-general forms is evolved and this is applied to the problem of obtaining the Cayley form of the intersection of two varieties from their Cayley forms. A solution is given based on the assumption of the existence of a suitable polynomial solution. It is shown that in some cases this assumption is true, while in others it is false. A necessary and sufficient identity for a form to be a Cayley form is also given.

This dissertation is not substantially the same as one which has been submitted at any other university. The abstract treatment of tensors given in section A is original and references are given where known results, other than elementary ones, are used or proved. The tensor treatment given of the questions considered in chapters B 1 and B 3 is new although many of the results are well-known and references are given. The essential results and techniques given in chapters B 2 and B 4 are, as far as I know, entirely original.

I wish to thank my supervisor, Dr J. A. Todd, for advice and encouragement in the preparation of this dissertation, and also Professor W. V. D. Hodge, who suggested the main problem considered in chapter B 4, which was the original motive for the techniques developed here.

I have also to thank D.S.I.R. for a grant.

September, 1956 Roger Penrose

CONTENTS

Section A : An algebraic theory of tensor systems

A 1 Abstract tensor systems ..... p. 1
A 2 Linear equations in a tensor system 60
A 3 Complete tensor systems 82
A 4 Tensors representing linear sets 94

Section B : Some applications of tensor techniques to algebraic geometry

B 1 Some general theorems 109
B 2 Non-general forms 126
B 3 Miscellaneous applications 148
B 4 Cayley forms 171

References 219

Section A : An algebraic theory of tensor systems.

An abstract tensor system will be defined as a set of algebraic entities which is closed with respect to operations analogous to those of tensor addition, (outer) multiplication, contraction and permutation of indices. The relationship of such a system to a system of ordinary tensors is analogous to that of a general ring to, say, a ring of $n \times n$ matrices. The properties of tensor systems considered here will mainly be in connection with the geometrical applications of section B. More generally they are intended as contributions to the philosophy of regarding a tensor (or a matrix) as an algebraic entity in its own right, its representation as a set of components being only of secondary importance.

A 1 Abstract tensor systems.

In this chapter an abstract definition of a tensor system will be given. However, my object is not primarily to obtain a strict logical development but simply to show that such an abstract definition can, in fact, be made. A large part of the treatment will be concerned with obtaining a consistent and a manageable notation.

Certain more restricted types of tensor systems will be discussed and possible identities of a particular kind will be investigated.

A 1.1 The tensor operations

Let $a^{i_1 \cdots i_r}_{j_1 \cdots j_s}$, $b^{k_1 \cdots k_t}_{l_1 \cdots l_u}$ be ordinary tensors, i.e. sets of numbers (or functions) which are subject to certain transformation laws. Then, in accordance with the laws of tensor algebra (see, for example, Eisenhart (5) pp.12-14 or Hodge (7) pp 17-22), the sum

$$a^{i_1 \cdots i_r}_{j_1 \cdots j_s} + b^{i_1 \cdots i_r}_{j_1 \cdots j_s} = c^{i_1 \cdots i_r}_{j_1 \cdots j_s} \qquad \text{A 1.1a}$$

can be formed, provided $r = t$ and $s = u$. The set of numbers $c^{i_1 \cdots i_r}_{j_1 \cdots j_s}$ then also constitutes a tensor. Also, the (outer) product

$$a^{i_1 \cdots i_r}_{j_1 \cdots j_s} \; b^{k_1 \cdots k_t}_{l_1 \cdots l_u} = d^{i_1 \cdots i_r \, k_1 \cdots k_t}_{j_1 \cdots j_s \, l_1 \cdots l_u} \qquad \text{A 1.1b}$$

can be formed for any two tensors and the set $d^{i_1 \cdots \; k_t}_{j_1 \cdots \; l_u}$ constitutes a tensor. Furthermore, if $r \geqslant 1 \leqslant s$, the operations of contraction

$$\sum_m a^{i_1 \cdots i_p \, m \, i_{p+2} \cdots i_r}_{j_1 \cdots j_q \, m \, j_{q+2} \cdots j_s} = e^{i_1 \cdots i_p \, i_{p+2} \cdots i_r}_{j_1 \cdots j_q \, j_{q+2} \cdots j_s} \qquad \text{A 1.1c}$$

may be performed (involving one upper and one lower index) to obtain new tensors. (The summation sign is usually omitted.) The set of such tensors is therefore closed with respect to these operations of addition, multiplication and contraction.

There is another operation which can be applied to tensors to produce new tensors and which is not usually stated explicitly. This is

simply the operation of permutation of indices. Thus, if the set of numbers $f^{i_1 \cdots i_r}_{j_1 \cdots j_s}$ is defined by

$$a^{i_{g(1)} \cdots i_{g(r)}}_{j_{h(1)} \cdots j_{h(s)}} = f^{i_1 \cdots i_r}_{j_1 \cdots j_s} \qquad \text{A 1.1d}$$

where g and h denote permutations of (1, .., r) and (1, .., s) respectively, then $f^{i_1 \cdots i_r}_{j_1 \cdots j_s}$ is also a tensor. I shall call this operation <u>transposition</u> and the tensor $f^{i_1 \cdots i_r}_{j_1 \cdots j_s}$ a <u>transpose</u> of $a^{i_1 \cdots i_r}_{j_1 \cdots j_s}$. The operation is analogous to that of forming the transpose of a matrix.

There is a certain notational difficulty here. For if the set of numbers a_{ij} constitutes the tensor a, the transpose a' of this tensor is given by the set of numbers a'_{ij} where $a'_{ij} = a_{ji}$. If the tensor a is represented by the symbol a_{ij}, it is not in order to represent the tensor a' by a_{ji}, there being no reason to consider the index i first, rather than j. (The tensors a and a' may, of course, be quite distinct, i.e. $a - a' \neq 0$.)

To avoid this difficulty, it will be necessary to assign a definite order to the upper indices and to the lower indices of a tensor. This order may be indicated by numbering the upper indices with consecutive integers and similarly for the lower indices. Thus, the tensor represented by the symbol $a^{i_2 i_1 i_3 \cdots i_r}_{j_1 \cdots j_{a-2} j_a j_{a-1}}$, for example, is a transpose of $a^{i_1 \cdots i_r}_{j_1 \cdots j_s}$. Similarly, the product of the tensor $a^{i_1 \cdots i_r}_{j_1 \cdots j_s}$ with the tensor $b^{i_1 \cdots i_t}_{j_1 \cdots j_u}$ is the tensor $a^{i_1 \cdots i_r}_{j_1 \cdots j_s} b^{i_{r+1} \cdots i_{r+t}}_{j_{s+1} \cdots j_{s+u}}$.

Tensors possessing indices of different types, that is, indices corresponding to different fields of transformation, may also be considered. The upper and lower indices of the tensor $a^{i_1 \ldots i_m}_{j_1 \ldots j_n}$ are of different types (corresponding to contragredient and cogredient fields of transformation). I shall call them conjugate index types. More generally, a tensor may possess indices of several quite distinct types for which the range of possible values taken by the indices may be different in the different cases. For example, the tensor a_{ij} may be a set of numbers given by $i = 1, .., n$ but $j = 1,..,m$ with $m \neq n$ (i.e. constituting a rectangular matrix), the field of transformation corresponding to the index i being different from that corresponding to j. Thus a_{ij} might be the tensor corresponding to a bilinear form $\sum_{i,j} a_{ij}x^i y^j$ where the fields of transformation of the two sets of variables x^i and y^j are distinct.

The rules A 1.1a, ..,A 1.1d for forming new tensors out of a given set of tensors must now be slightly modified so that tensors with indices of different types may be considered. The sum A 1.1a is allowed, provided that the two tensors concerned possess the same number of indices of each type. The product A 1.1b is always allowed. Contraction A 1.1c is permitted only between indices of conjugate types and transposition A 1.1d can only occur between indices of the same type. In view of this last remark, it is not necessary to number all the indices

of a tensor with distinct integers ; the indices corresponding to each index type must simply be ordered among themselves.

It will often be convenient to group together a number of tensor indices of various types and to consider this group of indices as a new tensor index, whose type is defined (formally) as the product of the original index types involved. The rules A 1.1a, ..., A 1.1d will then also apply to such composite index types. The index types that can be built up in this way form a free commutative semigroup $\mathfrak{S}$, whose generators are the original distinct types of tensor index considered. Any tensor is associated with a unique element of $\mathfrak{S}$, namely that index type which is the product of all the index types of the individual indices of the tensor. This element of $\mathfrak{S}$ will be called the <u>type</u> of the <u>tensor</u>. The semigroup $\mathfrak{S}$ is supposed to possess an identity element 1. The scalars (invariants) are then the tensors of type 1. The type of the product of two tensors is the product of their types. Each element of $\mathfrak{S}$ has a conjugate element in $\mathfrak{S}$ (the conjugate index type). The conjugate of the product of two elements of $\mathfrak{S}$ is the product of their conjugates.

A 1.2 <u>Abstract tensor systems</u> : definition

Let $\mathfrak{S}$ be a free commutative semigroup (containing an identity 1) with generators $\rho, \rho^*, \sigma, \sigma^*, \cdots$ (not necessarily finite in

number). The elements of $\mathfrak{S}$ are of the form $\rho^{r}\rho^{*r'}\sigma^{s}\sigma^{*s'}\ldots$ where r, r', s, s', ... are non-negative integers, of which only a finite number are non-zero. (The usual convention about zero exponents is being used here, e.g. $\rho^{0}=1$.) Two elements $\rho^{r_1}\rho^{*r_1'}\sigma^{s_1}\sigma^{*s_1'}\ldots$ and $\rho^{r_2}\rho^{*r_2'}\sigma^{s_2}\sigma^{*s_2'}\ldots$ of $\mathfrak{S}$ are equal only if $r_1 = r_2$, $r_1' = r_2'$, $s_1 = s_2$, ... Their product is $\rho^{r_1+r_2}\rho^{*(r_1'+r_2')}\sigma^{s_1+s_2}\ldots$. The semigroup $\mathfrak{S}$ possesses an involution $\alpha \to \alpha^*$. The element α^* is called the <u>conjugate</u> of α and is defined by $(\rho^{r}\rho^{*r'}\sigma^{s}\ldots)^* = \rho^{r'}\rho^{*r}\sigma^{s'}\ldots$. We have $(\alpha^*)^* = \alpha$ and $(\alpha\beta)^* = \alpha^*\beta^*$.

The set of elements $\rho, \sigma, \ldots$ will be called a set of <u>semigenerators</u> of $\mathfrak{S}$ if $\rho, \rho^*, \sigma, \sigma^*, \ldots$ are the generators of $\mathfrak{S}$. (I am only considering cases in which the semigenerators $\rho, \sigma, \ldots$ of $\mathfrak{S}$ are all distinct from their conjugates. Other cases could be considered, e.g. Cartesian tensors, for which no distinction is made between covariant and contravariant indices.)

An abstract tensor system $\mathfrak{J}$ over $\mathfrak{S}$ can now be defined. The definitions of the algebraic operations admitted by $\mathfrak{J}$ will be very closely bound up with the notation for the elements of $\mathfrak{J}$. It will be convenient to introduce this notation as part of the definition of some of the rules that these operations satisfy. The definition of an (abstract) <u>tensor system</u> $\mathfrak{J}$ over $\mathfrak{S}$ is now as follows :

A 1.2a $\mathcal{S}$ is the union of disjoint sets $\mathcal{S}_\alpha$, $\mathcal{S}_\beta$, ... where $\alpha, \beta, \ldots$ are the elements of $\mathfrak{G}$.

If $a \in \mathcal{S}_\alpha$ then a will be called an element of $\mathcal{S}$ of <u>type</u> α.

A 1.2b <u>Addition</u>. Each set $\mathcal{S}_\alpha$ is an additive abelian group with zero 0_α (or simply 0). That is, $\mathcal{S}_\alpha$ admits a binary operation + and a unary operation −, such that for all $a, b, c \in \mathcal{S}_\alpha$,

$a + b = b + a$, $a + (b + c) = (a + b) + c$, $a + 0_\alpha = a$,

$a + (-a) = 0_\alpha$.

The sums $a + (b + c)$ and $a + (-b)$ are usually written $a + b + c$ and $a - b$, respectively ; etc.

Addition is not defined between elements of distinct sets $\mathcal{S}_\alpha$.

A 1.2c <u>Multiplication</u>. The set $\mathcal{S}$ is a (non-commutative) semigroup with respect to a binary operation denoted simply by juxtaposition of symbols. For all $a, b, c \in \mathcal{S}$, $a(bc) = (ab)c$ this being usually written abc.

This operation will sometimes be called <u>outer product</u>.

A 1.2d The two distributive laws

$$a(b + c) = ab + ac, \quad (b + c)a = ba + ca$$

hold, for all $a, b, c \in \mathcal{S}$ for which the sum $b + c$ exists.

A 1.2e If $a \in \mathcal{S}_\alpha$ and $b \in \mathcal{S}_\beta$ then $ab \in \mathcal{S}_{\alpha\beta}$.

(Thus, $\mathcal{S}$ can be regarded as a kind of graded ring.)

CHAPTER 6

A 1.2f If $a \in \mathfrak{S}_\alpha$, $b \in \mathfrak{S}_\beta$ and $ab = 0_{\alpha\beta}$, then either $a = 0_\alpha$ or $b = 0_\beta$.

The converse result that $0_\alpha b = 0_{\alpha\beta} = a0_\beta$, follows from A 1.2b, A 1.2d and A 1.2e.

A 1.2g Notation. Let $\alpha = \rho^\lambda \rho^{*\lambda'} \sigma^\mu \sigma^{*\mu'} \ldots \in \mathfrak{G}$, where $\rho, \sigma, \ldots$ are semigenerators of $\mathfrak{G}$. Then if a is an element of $\mathfrak{S}_\alpha$, the notation

$$a = a_{\rho_1 \ldots \rho_\lambda \rho_1^* \ldots \rho_{\lambda'}^* \sigma_1 \ldots \sigma_\mu \sigma_1^* \ldots \sigma_{\mu'}^* \ldots}$$

may be used. If the indices $\rho_1, \ldots, \sigma_{\mu'}^*, \ldots$ are permuted in such a way that the order of $\rho_1, \ldots, \rho_\lambda$ remains unaltered and so also does that of $\rho_1^*, \ldots, \rho_{\lambda'}^*$, of $\sigma_1, \ldots, \sigma_\mu$ and of $\sigma_1^*, \ldots, \sigma_{\mu'}^*$, etc., then the symbol so obtained represents the same element a of $\mathfrak{S}_\alpha$. If the indices are permuted in any manner whatever, then the symbol still represents an element of $\mathfrak{S}_\alpha$, but it may be an element different from a. Thus, for example,

$$a = a_{\sigma_1 \rho_1^* \ldots \rho_{\lambda'-1}^* \rho_1 \rho_{\lambda'}^* \sigma_2 \rho_2 \ldots \rho_\lambda \sigma_3 \ldots \sigma_\mu \sigma_1^* \ldots \sigma_{\mu'}^* \ldots}$$

but the element

$$a_{\rho_2 \rho_1 \rho_3 \ldots \rho_\lambda \rho_1^* \ldots \rho_{\lambda'}^* \sigma_1 \ldots \sigma_\mu \ldots}$$

will generally be distinct from a.

(Such a set of suffixes can be regarded as an operator operating on the element a to produce a new element of $\mathfrak{S}_\alpha$. It

is an identity operator if the suffixes are numbered in their natural order.)

A 1.2h Transposition. If $a \in \mathfrak{S}_\alpha$ as in A 1.2g, so that a may be represented by the symbol

$$a_{\rho_1 \cdots \rho_\lambda \rho_1^* \cdots \rho_\lambda^* \sigma_1 \cdots \sigma_\alpha \cdots}$$

Then, according to A 1.2h, any symbol which is obtained from this by a permutation π of the indices, will represent some element b of $\mathfrak{S}_\alpha$, possibly distinct from a. This may be written

$$b = b_{\rho_1 \cdots \rho_\lambda \rho_1^* \cdots \rho_\lambda^* \sigma_1 \cdots \sigma_\alpha \cdots}$$

$$= (a)_\pi = a_{\pi(\rho_1) \cdots \pi(\rho_\lambda) \pi(\rho_1^*) \cdots \pi(\sigma_1) \cdots}$$

The element $b = (a)_\pi$ will be called a transpose of a and the permutation π will be called an α-transposition.

A 1.2i Let $a \in \mathfrak{S}_\alpha$, let π and ω be α-transpositions and let $b = (a)_\pi$. Let $\pi \cdot \omega$ denote the permutation π followed by ω Then $(b)_\omega = (a)_{\pi \cdot \omega}$, so that

$$((a)_\pi)_\omega = a_{\omega(\pi(\rho_1)) \cdots \omega(\pi(\rho_\lambda)) \omega(\pi(\rho_1^*)) \cdots}$$

A 1.2j As a special case of A 1.2i, we deduce that if π^{-1} is the permutation inverse to π and $b = (a)_\pi$, then $a = (b)_{\pi^{-1}}$, so that a is a transpose of each of its transposes.

A 1.2k It follows from A 1.2j and A 1.2g that if the indices of

$$a_{\pi(\rho_1) \cdots \pi(\rho_\lambda) \pi(\rho_1^*) \cdots \pi(\sigma_1) \cdots}$$

CHAPTER 6

are permuted, the symbol so obtained will represent the same element $(a)_\pi$ of $\mathfrak{F}_\alpha$, if the orders in which the sets of indices $\rho_1, \cdots, \rho_\lambda$; $\rho^*_1, \cdots, \rho^*_{\lambda'}$; $\sigma_1, \cdots, \sigma_\delta$; $\cdots$ appear are each unaltered. Hence a has at most $\lambda!\,\lambda'!\,\delta!\,\delta'! \cdots$ transposes.

A 1.21 If $a, b \in \mathfrak{F}_\alpha$ and π is an α-transposition, then

$$(a + b)_\pi = (a)_\pi + (b)_\pi$$

A 1.2m Notation. Let $a = a_{\rho_1 \cdots \rho_\lambda \rho^*_1 \cdots \rho^*_{\lambda'} \sigma_1 \cdots \sigma_\delta \cdots} \in \mathfrak{F}_{\rho^\lambda \rho^{*\lambda'} \sigma^\delta \cdots}$ and $b = b_{\rho_1 \cdots \rho_\mu \rho^*_1 \cdots \rho^*_{\mu'} \sigma_1 \cdots \sigma_\nu \cdots} \in \mathfrak{F}_{\rho^\mu \rho^{*\mu'} \sigma^\nu \cdots}$

Then the product of a with b will be written

$$ab = a_{\rho_1 \cdots \rho_\lambda \rho^*_1 \cdots \rho^*_{\lambda'} \sigma_1 \cdots \sigma_\delta \cdots} b_{\rho_{\lambda+1} \cdots \rho_{\lambda+\mu} \rho^*_{\lambda'+1} \cdots \rho^*_{\lambda'+\mu'} \sigma_{\delta+1} \cdots \sigma_{\delta+\nu} \cdots}$$

That is, for each index type $\rho, \rho^*, \sigma, \cdots$, the indices of a are numbered first and those of b afterwards. The transposes of ab are then obtained by applying permutations to the indices, which do not alter the number of indices of each type possessed by each of a and b. The symbol a....b.... is a compound symbol, whose indices obey the same rules (A 1.2g) with respect to permutations as does a symbol c...., except that the number of indices of each type possessed by a and by b must not be changed.

A 1.2n Notation. Let π be an α-transposition and let ω be a β-transposition ; $\alpha = \rho^\lambda \rho^{*\lambda'} \sigma^\delta \cdots$, $\beta = \rho^\mu \rho^{*\mu'} \sigma^\nu \cdots$. The

$\alpha\beta$ -transposition $\pi\omega$ is then defined to be that permutation of the indices $\rho_1 \cdots \rho_\lambda \rho_1^* \cdots \rho_\lambda^* \sigma_1 \cdots \sigma_a \cdots \rho_{\lambda+1} \cdots \rho_{\lambda+\mu} \rho_{\lambda+1}^* \cdots \rho_{\lambda+\mu}^* \sigma_{a+1} \cdots \sigma_{a+\nu}$ which is obtained by performing the permutation π on the first set $(\rho_1 \cdots \sigma_a \cdots)$ and the appropriate permutation corresponding to ω on the second set $(\rho_{\lambda+1} \cdots \sigma_{a+\nu} \cdots)$. For example if $\alpha = \rho^2 \sigma^2 \sigma^*$ and π sends $\rho_1 \rho_2 \sigma_1 \sigma_2 \sigma_1^*$ into $\rho_2 \rho_1 \sigma_1 \sigma_2 \sigma_1^*$, while $\beta = \rho^3 \rho^* \sigma^3$ and ω sends $\rho_1 \rho_2 \rho_3 \rho_1^* \sigma_1 \sigma_2 \sigma_3$ into $\rho_3 \rho_2 \rho_1 \rho_1^* \sigma_2 \sigma_3 \sigma_1$, then $\pi\omega$ sends $\rho_1 \rho_2 \sigma_1 \sigma_2 \sigma_1^* \rho_3 \rho_4 \rho_5 \rho_1^* \sigma_3 \sigma_4 \sigma_5$ into $\rho_2 \rho_1 \sigma_1 \sigma_2 \sigma_1^* \rho_5 \rho_4 \rho_3 \rho_1^* \sigma_4 \sigma_5 \sigma_3$; that is, it sends $\rho_1 \cdots \rho_5 \rho_1^* \sigma_1 \cdots \sigma_5 \sigma_1^*$ into $\rho_2 \rho_1 \rho_5 \rho_4 \rho_3 \rho^* \sigma_1 \sigma_2 \sigma_4 \sigma_5 \sigma_3 \sigma_1^*$.

In general, $\pi\omega \neq \omega\pi$. However, for any three transpositions π, ω, ξ we have $(\pi\omega)\xi = \pi(\omega\xi)$. Also, if π and ω are both α -transpositions while ξ and η are both β -transpositions, we have $(\pi \cdot \omega)(\xi \cdot \eta) = (\pi\xi) \cdot (\omega\eta)$ (see A 1.2i). These relations may be compared with A 1.5g.

A 1.2o If π is an α-transposition, ω is a β -transposition, $a \in \mathfrak{S}_\alpha$ and $b \in \mathfrak{S}_\beta$, then

$$(a)_\pi (b)_\omega = (ab)_{\pi\omega}$$

A 1.2p If $\mathfrak{a} \in \mathcal{J}_{\rho^\lambda \rho^{*\lambda'} \sigma^\alpha \ldots}$ and $\mathfrak{b} \in \mathcal{J}_{\rho^\mu \rho^{*\mu'} \sigma^\nu \ldots}$,
then

$$\mathfrak{a}_{\rho_1 \ldots \rho_\lambda \rho^*_1 \ldots \rho^*_{\lambda'} \sigma_1 \ldots \sigma_\alpha \ldots} \quad \mathfrak{b}_{\rho_{\lambda+1} \ldots \rho_{\lambda+\mu} \rho^*_{\lambda'+1} \ldots \rho^*_{\lambda'+\mu'} \sigma_{\alpha+1} \ldots \sigma_{\alpha+\nu} \ldots}$$

$$= \mathfrak{b}_{\rho_{\lambda+1} \ldots \rho_{\lambda+\mu} \rho^*_{\lambda'+1} \ldots \rho^*_{\lambda'+\mu'} \sigma_{\alpha+1} \ldots \sigma_{\alpha+\nu} \ldots} \mathfrak{a}_{\rho_1 \ldots \rho_\lambda \rho^*_1 \ldots \rho^*_{\lambda'} \sigma_1 \ldots \sigma_\alpha \ldots}$$

Thus, ab is a transpose of ba (see A 1.2m).

Systems in which this law holds could be called commutative tensor systems. Non-commutative tensors could also be studied (e.g tensors with, say, quaternion components) but I shall not be concerned with them here.

A 1.2q Contraction. Let τ be any generator of $\mathfrak{S}$. Then the operation of τ-contraction is defined as a mapping of the set of elements of $\mathcal{J}_{\alpha\tau\tau^*}$ into $\mathcal{J}_\alpha$, for each $\alpha \in \mathfrak{S}$. The operation of τ^*-contraction is the same as that of τ-contraction. The τ-contraction of the element a will be written $(\mathfrak{a})_{[\tau]}$ (or $(\mathfrak{a})_{[\tau^*]}$). If $\mathfrak{a} = \mathfrak{a}_{\rho_1 \ldots \rho_\lambda \ldots \tau_1 \ldots \tau_t \tau^*_1 \ldots \tau^*_{t'}}$ ($\mathfrak{a} \in \mathcal{J}_{\rho^\lambda \ldots \tau^t \tau^{*t'} \ldots}$; $t, t' > 0$), then I write

$$(\mathfrak{a})_{[\tau]} = (\mathfrak{a})_{[\tau]\, \rho_1 \ldots \rho_\lambda \ldots \tau_1 \ldots \tau_{t-1} \tau^*_1 \ldots \tau^*_{t'-1} \ldots}$$

$$= \mathfrak{a}_{\rho_1 \ldots \rho_\lambda \ldots \tau_1 \ldots \tau_{t-1} \overline{\tau}_1 \tau^*_1 \ldots \tau^*_{t'-1} \overline{\tau}^*_1 \ldots}$$

Thus, to form the τ-contraction of a, the indices τ_t and $\tau^*_{t'}$ of a are replaced by the dummy indices $\overline{\tau}_1$ and $\overline{\tau}^*_1$, respectively, which do not contribute to the type of a. In the

same way, if $b = (a)_{\pi}$ is a transpose of a, the τ-contraction of b may be written as the symbol a with permuted indices and with the symbols $\overline{\tau}_i$ and $\overline{\tau}_i^*$ replacing τ_t and τ_t^*, respectively.

As an example, if $a \in \mathcal{S}_{\rho^2\sigma^2\sigma^*\tau^3\tau^{*2}}$ and

$$b = a_{\rho_2\rho_1\sigma_1\sigma_2\sigma_1^*\tau_2\tau_3\tau_1\tau_2^*\tau_1^*},$$

then the τ-contraction of b is given by

$$(b)_{[\tau]} = b_{\rho_1\rho_2\sigma_1\sigma_2\sigma_1^*\tau_1\tau_2\overline{\tau}_1\tau_1^*\overline{\tau}_1^*}$$

$$= a_{\rho_2\rho_1\sigma_1\sigma_2\sigma_1^*\tau_2\overline{\tau}_1\tau_1\overline{\tau}_1^*\tau_1^*}$$

while the σ-contraction of b is given by

$$(b)_{[\sigma]} = a_{\rho_2\rho_1\sigma_1\overline{\sigma}_1\overline{\sigma}_1^*\tau_2\tau_3\tau_1\tau_2^*\tau_1^*}$$

If a symbol representing an element of some $\mathcal{S}_{\alpha\tau\tau^*}$ already possesses the dummy indices $\overline{\tau}_1$ and $\overline{\tau}_1^*$, then its τ-contraction may be denoted by the same symbol but with the highest numbered indices τ_i and τ_j replaced by $\overline{\tau}_2$ and $\overline{\tau}_2^*$, respectively. Similarly, if $\overline{\tau}_1$, $\overline{\tau}_2$, $\overline{\tau}_1^*$ and $\overline{\tau}_2^*$ already appear, then $\overline{\tau}_3$ and $\overline{\tau}_3^*$ may be used; and so on. For example, the τ-contraction of the element $(b)_{[\tau]}$ above may be written as

$$((b)_{[\tau]})_{[\tau]} = a_{\rho_2\rho_1\sigma_1\sigma_2\sigma_1^*\overline{\tau}_2\overline{\tau}_1\tau_1\overline{\tau}_1^*\overline{\tau}_2^*}.$$

If a symbol representing an element of some $\mathcal{S}_{\alpha\tau\tau^*}$ possesses dummy indices corresponding to a semigenerator of $\mathfrak{S}$ distinct from τ or τ^*

and perhaps also dummy indices corresponding to τ and τ^*, then the τ-contraction of this element may be obtained in the same way. That is, the highest numbered τ_i and τ_j^* may be replaced by $\overline{\tau}_k$ and $\overline{\tau}_k^*$, respectively, where $\overline{\tau}_{k-1}$ and $\overline{\tau}_{k-1}^*$ are the highest numbered $\overline{\tau}_i$ and $\overline{\tau}_i^*$ already appearing. The other indices are left unchanged.

Implicit in this last notation is the following rule to ensure its consistency :

A 1.2r Let τ and σ be generators of $\mathfrak{S}$ and let $\alpha \in \mathfrak{S}$. Then for each $a \in \mathcal{J}_{\alpha\sigma\sigma^*\tau\tau^*}$,

$$((a)_{[\tau]})_{[\sigma]} = ((a)_{[\sigma]})_{[\tau]}$$

Thus, for example, if b is as in A 1.2q, the τ-contraction of $(b)_{[\sigma]}$ and the σ-contraction of $(b)_{[\tau]}$ may both be written as

$$a_{\rho_2\rho_1\sigma_1\overline{\sigma}_1\overline{\sigma}_1^*\tau_2\overline{\tau}_1\tau_1\overline{\tau}_1^*\tau_1^*}$$

(The rule A 1.2r is trivial if $\sigma = \tau$ or $\sigma = \tau^*$.)

A 1.2s Let τ be a generator of $\mathfrak{S}$ and let $\alpha \in \mathfrak{S}$. Let $a = a_{\rho_1\cdots\rho_n\cdots\tau_1\cdots\tau_1^*\cdots} \in \mathcal{J}_{\alpha\tau\tau^*}$ and π be a permutation of the indices of a which leaves the highest numbered τ_i and τ_j^* unaltered. Then

$$((a)_\pi)_{[\tau]} = ((a)_{[\tau]})_\pi$$

Hence, if $b = b_{\sigma_1\sigma_2\tau_1\tau_2\tau_3\tau_4\tau_1^*\tau_2^*\tau_3^*} \in \mathcal{J}_{\sigma^2\tau^4\tau^{*3}}$ for example, and c is the transpose of b given by

$$c = b_{\sigma_2\sigma_1\tau_2\tau_3\tau_1\tau_4\tau_2^*\tau_1^*\tau_3^*}$$

then

$$(c)_{[\tau]} = (b)_{[\tau]\,\sigma_2\sigma_1\tau_2\tau_3\tau_1\tau_2^*\tau_1^*}$$
$$(= b_{\sigma_2\sigma_1\tau_2\tau_3\tau_1\bar{\tau}_1\tau_2^*\tau_1^*\bar{\tau}_1^*})$$

Thus, the transposes of $(a)_{[\tau]}$ are obtained by permuting the indices, of each type, of $a_{\rho_1\cdots\rho_\lambda\ \cdots\cdots\ \tau_1\cdots\tau_{k-1}\bar{\tau}_1\tau_1^*\cdots\tau_{k'-1}^*\bar{\tau}_1^*\ \cdots\cdots}$ among themselves, but leaving the positions of $\bar{\tau}_1$ and $\bar{\tau}_1^*$ unaltered. Similarly, it follows from this and A 1.2i that the transposes of the τ-contraction of any transpose $(a)_\omega$ of a can be obtained by permuting the indices of each type, except $\bar{\tau}_1$ and $\bar{\tau}_1^*$, among themselves, of that symbol representing $(a)_\omega$ which is obtained (see A 1.2q) by a suitable permutation of the indices of $a_{\rho_1\ \cdots\cdots\ \tau_{k-1}\bar{\tau}_1\tau_1^*\cdots\tau_{k'-1}^*\bar{\tau}_1^*\ \cdots\cdots}$.

Hence, in particular, if $b(\in \mathfrak{J}\sigma^2\tau^4\tau^{*3})$ is as above and

$$d = d_{\sigma_1\sigma_2\tau_1\tau_2\tau_3\tau_1^*\tau_2^*} = b_{\sigma_2\sigma_1\tau_2\bar{\tau}_1\tau_3\tau_1\bar{\tau}_1^*\tau_2^*\tau_1^*},$$

then

$$b_{\sigma_1\sigma_2\tau_3\bar{\tau}_1\tau_2\tau_1\bar{\tau}_1^*\tau_2^*\tau_1^*} = d_{\sigma_2\sigma_1\tau_1\tau_3\tau_2\tau_1^*\tau_2^*}$$

gives a transpose of d.

This same rule holds for symbols with more than one pair of dummy indices, since it holds for each pair in turn. Thus if

$e \in \mathfrak{J}\rho^2\sigma^3\sigma^*\tau^4\tau^{*2}$ and

$$f = e_{\rho_2\rho_1\sigma_1\bar{\sigma}_1\sigma_2\sigma_1^*\bar{\tau}_2\tau_1\tau_2\bar{\tau}_1\bar{\tau}_1^*\tau_2^*}$$

CHAPTER 6

then

$$e_{\rho_1\rho_2\sigma_2\bar{\sigma}\,\sigma_1\sigma_1^*\,\bar{\tau}_2\tau_2\tau_1\bar{\tau}_1\bar{\tau}_1^*\,\bar{\tau}_2^*} = f_{\rho_2\rho_1\sigma_2\sigma_1\tau_2\tau_1}$$

gives a transpose of f.

Rule A 1.2s enables us to regard a symbol possessing dummy indices as a compound symbol representing an element of $\mathcal{S}$, whose transposes are obtained by permuting the indices of the symbol of each type, apart from the dummy indices, among themselves. The notation introduced in A 1.2q is really only satisfactory with regard to the operation of transposition, in virtue of A 1.2s. Similarly, the following rule ensures that the notation is satisfactory in connection with the notation A 1.2m for multiplication.

A 1.2t Let $a \in \mathcal{S}_\alpha$ and $b \in \mathcal{S}_{\beta\tau\tau^*}$, where τ is a semigenerator of $\mathfrak{S}$. Then

$$a(b)_{[\tau]} = (ab)_{[\tau]}$$

Thus, for example, if $c \in \mathcal{S}_{\rho\tau^2\tau^*}$ and $d \in \mathcal{S}_{\rho\sigma\tau^2\tau^*}$, the symbol

$$c_{\rho_1\tau_1\tau_2\tau_1^*}\ d_{\rho_2\sigma_1\tau_2\tau_3\bar{\tau}_1\bar{\tau}_1^*}$$

denotes both the τ-contraction of cd and the product of c with the τ-contraction of d. The symbol

$$c_{\rho_1\tau_1\bar{\tau}_1\bar{\tau}_1^*}\ d_{\rho_2\sigma_1\tau_2\tau_3\tau_1^*}$$

denotes the product of $(c)_{[\tau]}$ with d, which is also the τ-contraction of the transpose

$$c_{\rho_1 \tau_1 \tau_4 \tau_2^*} \quad d_{\rho_2 \sigma_1 \tau_2 \tau_3 \tau_1^*}$$

cd. This follows from A 1.2p, A 1.2s and A 1.2t since

$$(c_{\rho_1 \tau_1 \tau_4 \tau_2^*} \quad d_{\rho_2 \sigma_1 \tau_2 \tau_3 \tau_1^*})_{[\tau]} = (d_{\rho_2 \sigma_1 \tau_2 \tau_3 \tau_1^*} \quad c_{\rho_1 \tau_1 \tau_4 \tau_2^*})_{[\tau]}$$

$$((dc)_{[\tau]})_{\rho_2 \rho_1 \sigma_1 \tau_2 \tau_3 \tau_1 \tau_1^*} = (d(c)_{[\tau]})_{\rho_2 \rho_1 \sigma_1 \tau_2 \tau_3 \tau_1 \tau_1^*} = (c)_{[\tau]} d .$$

The notation is therefore consistent in this respect.

A.1.2u If τ is a generator of $\mathfrak{G}$ and a, b $\in \mathfrak{J}_{\alpha \tau \tau^*}$, then

$$(a + b)_{[\tau]} = (a)_{[\tau]} + (b)_{[\tau]}$$

A 1.2v Any transpose or contraction of a zero of $\mathfrak{J}$ is again a zero of $\mathfrak{J}$.

This follows from A 1.2e, A 1.2u and A 1.2b.

A 1.2w Let $\rho, \cdots, \tau, \cdots$ be a set of semigenerators of $\mathfrak{G}$ and let

$$a = a_{\rho_1 \cdots \rho_\lambda \cdots \cdots \tau_1 \cdots \tau_k \tau_1^* \cdots \tau_{\lambda'}^* \cdots} \in \mathfrak{J}_{\rho^\lambda \cdots \tau^x \tau^{x'} \cdots}$$ Then

$$((a)_{[\tau]})_{[\tau]} = ((a_{\rho_1 \cdots \rho_\lambda \cdots \cdots \tau_1 \cdots \tau_{k-2} \tau_k \tau_{k+1} \tau_1^* \cdots \tau_{x'-2}^* \tau_{x'}^* \tau_{x'-1}^* \cdots})_{[\tau]})_{[\tau]}$$

That is,

$$a_{\rho_1 \cdots \tau_{k-2} \bar{\tau}_2 \bar{\tau}_1 \tau_1^* \cdots \tau_{x'-2}^* \bar{\tau}_2^* \bar{\tau}_1^* \cdots} = a_{\rho_1 \cdots \tau_{k-2} \bar{\tau}_1 \bar{\tau}_2 \tau_1^* \cdots \tau_{x'-2}^* \bar{\tau}_1^* \bar{\tau}_2^* \cdots} .$$

Thus, if a symbol representing an element of $\mathfrak{J}$ possesses the dummy indices $\bar{\tau}_1, \bar{\tau}_2, \bar{\tau}_1^*$ and $\bar{\tau}_2^*$, it will represent the same element of $\mathfrak{J}$ if the pairs $\bar{\tau}_1, \bar{\tau}_1^*$ and $\bar{\tau}_2, \bar{\tau}_2^*$ are interchanged. It follows from this that if a symbol possesses dummy indices $\bar{\tau}_k, \bar{\tau}_{k+1}, \bar{\tau}_k^*, \bar{\tau}_{k+1}^*$ the pairs $\bar{\tau}_k, \bar{\tau}_k^*$ and $\bar{\tau}_{k+1}, \bar{\tau}_{k+1}^*$ may also be interchanged without effect. Hence, if a sumbol possesses the dummy indices $\bar{\tau}_1, \cdots, \bar{\tau}_t, \bar{\tau}_1^*, \cdots, \bar{\tau}_t^*$ and a permutation is applied to the indices $\bar{\tau}_1, \cdots, \bar{\tau}_t$,

it still represents the same element of $\mathfrak{J}$ provided that this permutation is also applied to $\bar{\tau}_1^*, \ldots, \bar{\tau}_x^*$ and the other indices are left unaltered.

For example, if $b \in \mathfrak{J}_{\rho\,\tau^4\,\tau^{*3}}$, then

$$b_{\rho_1\bar{\tau}_2\tau_1\bar{\tau}_3\bar{\tau}_1\bar{\tau}_3^*\bar{\tau}_2^*\bar{\tau}_1^*} = b_{\rho_1\bar{\tau}_3\tau_1\bar{\tau}_1\bar{\tau}_2\bar{\tau}_1^*\bar{\tau}_3^*\bar{\tau}_2^*}$$

but this element will in general be distinct from the element $b_{\rho_1\bar{\tau}_3\tau_1\bar{\tau}_1\bar{\tau}_2\bar{\tau}_3^*\bar{\tau}_2^*\bar{\tau}_1^*}$, etc.

The rules for the notation introduced in A 1.2g, A 1.2m, A 1.2q can now be summed up. Let $\rho, \ldots, \tau, \ldots$ be a set of semigenerators of $\mathfrak{S}$ and let the symbol a denote an element of $\mathfrak{J}_{\rho^\lambda\rho^{*\lambda'}\ldots\tau^\kappa\tau^{*\kappa'}\ldots}$. Then, other elements of $\mathfrak{J}$ may be obtained from this symbol by supplying it with a set of suffixes $\rho_1, \rho_2, \ldots, \rho_{\lambda+\lambda''}, \bar{\rho}_1^*, \ldots, \bar{\rho}_{\lambda''}^*, \ldots, \tau_1, \ldots, \tau_{\kappa+\kappa''}, \bar{\tau}_1, \ldots, \bar{\tau}_{\kappa''}, \tau_1^*, \ldots$ in some order, for some $r'' \leq r, r'$; ..... ; $t'' \leq t, t'$. The suffixes, of each different type which appears, must be numbered with consecutive integers 1, 2, .., so that the suffixes are all distinguished from one another ; the letter ρ must appear $r + r'$ times in all, r' times with an asterisk and r times without, etc., ... ; the number of $\bar{\rho}$'s must be the same as the number of $\bar{\rho}^*$'s, etc., ... Similarly, if b, c, .. are also symbols denoting elements of $\mathfrak{J}$,

new elements of $\mathcal{S}$ can be built up by writing the symbol a with its suffixes followed by b with its suffixes (as a product), followed, perhaps, by c with its suffixes, etc. The numbering of the suffixes must take place throughout the whole product, so that no suffix is duplicated in the product. The type of the element of $\mathcal{S}$ represented by such a symbol is given by the product of all the unbarred index types which appear. The transposes of the element are given by permuting the unbarred suffixes among themselves ; the τ-contraction of the element is obtained by replacing the highest numbered τ_i and τ^*_j by τ_k and τ^*_k, respectively, τ_k and τ^*_k being the highest numbered $\overline{\tau}_i$ and $\overline{\tau}^*_j$ already appearing.

Permutations, which do not change the element of $\mathcal{S}$ represented by the symbol, are given (see A 1.2g and A 1.2q) by permutations of the suffixes of each of a, b, .. which do not alter the order of $\rho_1, \cdots, \rho_{\lambda\cdots\lambda''}, \overline{\rho}_1, \cdots, \overline{\rho}_{\lambda''}$ or of $\rho^*_1, \cdots, \rho^*_{\lambda'\cdots\lambda''}, \overline{\rho}^*_1, \cdots, \overline{\rho}^*_{\lambda''}$, etc. ; or by permutations (see A 1.2p) of the symbols a..., b..., c..., among themselves (together with suffixes) ; or by any permutation (see A 1.2w) of barred suffixes of a given type which is accompanied by the same permutation of the conjugate barred suffixes. For example, if $a \in \mathcal{S}_{\rho^2\rho^*\sigma\tau}$, $b \in \mathcal{S}_{\rho\sigma\sigma^*}$ and $c \in \mathcal{S}_{\rho^*\sigma^2\tau}$, then

$$a_{\sigma_3 \bar{\rho}_1 \sigma_1 \tau_2 \bar{\rho}_1^* \rho_1} \quad b_{\sigma_2 \bar{\sigma}_1^* \rho_2} \quad c_{\sigma_4 \bar{\rho}_2^* \tau_1 \bar{\sigma}_1 \sigma_1}$$

$$= a_{\bar{\rho}_1 \rho_1 \bar{\rho}_1^* \sigma_3 \tau_2} \quad b_{\bar{\rho}_2 \sigma_2 \bar{\sigma}_1^*} \quad c_{\bar{\rho}_2^* \sigma_4 \bar{\sigma}_1 \sigma_1}$$

$$= c_{\bar{\rho}_2^* \sigma_4 \bar{\sigma}_1 \sigma_1} \quad b_{\bar{\rho}_2 \sigma_2 \bar{\sigma}_1^*} \quad a_{\bar{\rho}_1 \rho_1 \bar{\rho}_1^* \sigma_3 \tau_2}$$

$$= c_{\bar{\rho}_1^* \sigma_4 \bar{\sigma}_1 \sigma_1} \quad b_{\bar{\rho}_1 \sigma_2 \bar{\sigma}_1^*} \quad a_{\bar{\rho}_2 \rho_1 \bar{\rho}_2^* \sigma_3 \tau_2} \quad \in \mathfrak{J}_{\rho \sigma^4 \tau^2} .$$

A 1.2x The set $\mathfrak{J}_1$ is a field with respect to the operations of addition and multiplication.

That $\mathfrak{J}_1$ is a commutative ring without divisors of zero, (an integral domain if $\mathfrak{J}_1$ possesses a unit) follows from A 1.2b, .., A 1.2f' and A 1.2p. The sets $\mathfrak{J}_\alpha$, for each $\alpha \in \mathfrak{S}$, are clearly vector spaces over $\mathfrak{J}_1$ (in virtue of A 1.2b, .., A 1.2e). The elements of $\mathfrak{J}_1$ will be called **scalars**.

A 1.2y The field $\mathfrak{J}_1$ is without characteristic.

Thus, the field $\mathfrak{J}_1$ contains a subfield isomorphic with the field of rational numbers. This subfield will frequently be identified with the rational field and the usual notation 0, 1, 2,.., $\frac{1}{2}$, .. will often be used for its elements.

A 1.3 **Some simplifying notations.**

In order to avoid having to write an excessive number of indices it will be convenient to introduce a number of additional notational conventions.

A 1.3a **Composite indices.** Let $\alpha, \beta, \ldots \in \mathfrak{S}$ (not necessarily

generators of $\mathfrak{S}$) and let $a \in \mathcal{J}_{\alpha^p \alpha^{*p'} \beta^q \ldots}$. Then the notation

$$a = a_{\alpha_1 \ldots \alpha_p \alpha_1^* \ldots \alpha_{p'}^* \beta_1 \ldots \beta_q \ldots}$$

may be used. As in A 1.3g, the indices may be permuted in any way which leaves the orders of each of the sets $\alpha_1, \ldots, \alpha_p$; $\alpha_1^*, \ldots, \alpha_{p'}^*$; $\beta_1, \ldots, \beta_q$; $\ldots$ unaltered and the symbol still represents the same element a of $\mathcal{J}$. If they are permuted in any other way, the symbol obtained represents a transpose of a. Thus, if the indices $\alpha, \alpha^*, \beta, \ldots$ are each numbered in ascending order, the element represented is a, and each index is a composite index standing for a set of elementary indices $\rho_\lambda \ldots \rho_{\lambda+s} \rho_{\lambda'}^* \ldots \rho_{\lambda'+s'}^* \sigma_t \ldots \sigma_{t+u} \ldots$, where $\rho, \rho^*, \sigma, \ldots$ are the generators of $\mathfrak{S}$. However, if the arrangement of the indices $\alpha_1, \alpha_2, \ldots, \alpha_1^*, \ldots, \beta_1, \ldots$ is such that it is obtained from an ascending order arraggement by some permutation of each of the sets $\alpha_1, \ldots, \alpha_p$; $\alpha_1^*, \ldots, \alpha_{p'}^*$; $\beta_1, \ldots, \beta_q$; $\ldots$, then the transpose of a represented is that obtained by performing the same permutation on the corresponding groups of elementary indices $\rho_1, \ldots$

For example, let $\alpha = \rho^2 \rho^* \sigma$, $\beta = \sigma^2 \tau = \gamma$ and let $a \in \mathcal{J}_{\alpha\alpha^*\beta^2\gamma}$ $(= \mathcal{J}_{\rho^2\rho^{*3}\sigma^7\sigma^*\tau^3}$). Then

$$a = a_{\alpha_1 \alpha_1^* \beta_1 \beta_2 \gamma_1} = a_{\rho_1 \rho_2 \rho_1^* \sigma_1 \rho_3 \rho_2^* \rho_3^* \sigma_1^* \sigma_2 \sigma_3 \tau_1 \sigma_4 \sigma_5 \tau_2 \sigma_6 \sigma_7 \tau_3}$$

$$= a_{\rho_1 \ldots \rho_3 \rho_1^* \ldots \rho_3^* \sigma_1 \ldots \sigma_7 \sigma_1^* \tau_1 \ldots \tau_3}$$

$$(= a_{\sigma_1 \sigma_2 \tau_1 \sigma_3 \sigma_4 \tau_2 \rho_1 \rho_1^* \rho_2^* \sigma_1^* \sigma_5 \sigma_6 \tau_3 \rho_2 \rho_3 \rho_3^* \sigma_7} = a_{\beta_1 \gamma_1 \alpha_1^* \beta_2 \alpha_1}),$$

so that

$$a_{\alpha_1 \alpha_1^* \beta_2 \beta_1 \gamma_1} = a_{\rho_1 \rho_2 \rho_1^* \sigma_1 \rho_3 \rho_2^* \rho_3^* \sigma_1^* \sigma_4 \sigma_5 \tau_2 \sigma_2 \sigma_3 \tau_1 \sigma_6 \sigma_7 \tau_3}$$

$$= a_{\rho_1 \dots \rho_3 \rho_1^* \dots \rho_3^* \ \sigma_1 \sigma_4 \sigma_5 \sigma_2 \sigma_3 \sigma_6 \sigma_7 \sigma_1^* \tau_2 \tau_1 \tau_3}$$

If $b \in \mathfrak{F}_{\beta' \gamma}$, then $b = b_{\beta_1 \gamma_1 \beta_2} = b_{\sigma_1 \sigma_2 \tau_1 \sigma_3 \sigma_4 \tau_2 \sigma_5 \sigma_6 \tau_3}$,

so that

$$b_{\beta_2 \gamma_1 \beta_1} = b_{\sigma_5 \sigma_6 \tau_3 \sigma_3 \sigma_4 \tau_2 \sigma_1 \sigma_2 \tau_1} = b_{\sigma_5 \sigma_6 \sigma_3 \sigma_4 \sigma_1 \sigma_2 \tau_3 \tau_2 \tau_1} .$$

The product of a with b may be written

$$a_{\alpha_1 \alpha_1^* \beta_1 \beta_2 \gamma_1} \quad b_{\beta_3 \gamma_2 \beta_4} .$$

(It should be noted, however, that this is in general distinct from $b_{\beta_3 \gamma_2 \beta_4} \ a_{\alpha_1 \alpha_1^* \beta_1 \beta_2 \gamma_1}$.)

A 1.3b If an index type, composite or otherwise, occurs only once in a symbol (e.g. if α_1 appears but not α_2), then the suffix 1 may be omitted. Thus, the symbols $a_{\alpha_1 \alpha_1^* \beta_1 \beta_2 \gamma_1}$ and $b_{\beta_1 \gamma_1 \beta_2}$ above may be written respectively as

$$a_{\alpha \alpha^* \beta_1 \beta_2 \gamma} \quad \text{and} \quad b_{\beta_1 \gamma \beta_2}$$

Their product ab may be written as $a_{\alpha \alpha^* \beta_1 \beta_2 \gamma_1} \quad b_{\beta_3 \gamma_2 \beta_4}$.

A 1.3c If $\alpha \in \mathfrak{S}$ and $r \geqslant 0$ then the symbol $\alpha^\wedge$ may be used as an index in place of $\alpha_1 \dots \alpha_r$. (If $r = 0$ it is a vacuous index.) Such a symbol will generally only be used if it occurs only once in an expression, so that it needs no numerical suffix.

A 1.3d <u>α-contraction</u>. Let $\alpha, \beta \in \mathfrak{S}$ and let $\alpha = \rho^{\wedge}\rho^{*\wedge'}\sigma^{\wedge}\sigma^{*\wedge'}\ldots$. If $a \in \mathcal{S}_{\beta\alpha\alpha^*}$, then the α-contraction $(a)_{[\alpha]}$ of a is defined to be that element of $\mathcal{S}_\beta$ which is obtained from a by performing the operation of ρ-contraction r + r' times, σ-contraction s + s' times, etc. It does not matter in which order these contractions are performed because of A 1.2r.

It is clear that α-contraction satisfies laws corresponding to those of A 1.2r, .., A 1.2w, as follows at once from A 1.2r, .., A 1.2w. Also, we have $((a)_{[\alpha]})_{[\beta]} = (a)_{[\alpha\beta]}$.

A 1.3e <u>Dummy composite indiees</u>. The notation of A 1.3a may be extended to include dummy indices $\bar{\alpha}_i, \bar{\alpha}_i^*, \bar{\beta}_i$, ... Each such dummy index corresponds to a group of dummy indices $\bar{\rho}_r \cdots \bar{\rho}_{r+s}\bar{\rho}_{r'}^* \cdots \bar{\rho}_{r'+s'}^*$. Dummy indices must occur in pairs, e.g. $\bar{\alpha}_i$, $\bar{\alpha}_i^*$ and each elementary index which constitutes $\bar{\alpha}_i$ must be paired off with the corresponding index constituting $\bar{\alpha}_i^*$.

For example, if $\alpha = \rho^2\rho^*$ and $a \in \mathcal{S}_{\rho^3\rho^{*3}}$ we have

$$a_{\bar{\alpha}\bar{\alpha}^*} = a_{\bar{\rho}_1\bar{\rho}_2\bar{\rho}_3^*\bar{\rho}_1^*\bar{\rho}_2^*\bar{\rho}_3} = a_{\bar{\rho}_1\bar{\rho}_2\bar{\rho}_3\bar{\rho}_3^*\bar{\rho}_1^*\bar{\rho}_2^*} .$$

(The convention A 1.3b, about omitting the suffix 1 of an index which occurs singly, also applies to dummy indices.) If $\beta = \rho^2 = \gamma$ and $b \in \mathcal{S}_{\rho^4\rho^{*4}}$, then

$$b_{\bar{\beta}\bar{\gamma}\bar{\gamma}^*\bar{\beta}^*} = b_{\bar{\beta}_1\bar{\beta}_2\bar{\beta}_2^*\bar{\beta}_1^*} = b_{\bar{\rho}_1\bar{\rho}_2\bar{\rho}_3\bar{\rho}_4\bar{\rho}_3^*\bar{\rho}_4^*\bar{\rho}_1^*\bar{\rho}_2^*}$$

etc., etc.

It should be noted that dummy indices of different types cannot always be permuted without effect, as is the case with elementary dummy indices of different types. Thus, if b is as above,

$$b_{\bar\rho_1\bar\rho_2\bar\rho_3\bar\rho_4\rho_1^*\rho_2^*\rho_3^*\rho_4^*} = b_{\bar\gamma\bar\beta\bar\gamma^*\bar\beta^*} \neq b_{\bar\beta\bar\gamma\bar\gamma^*\bar\beta^*}$$

in general. Furthermore, the α-contraction of an element cannot always be obtained by simply replacing the highest numbered α_i and α_j^*, which occur, by $\bar\alpha_k$ and $\bar\alpha_k^*$. For example, if a is as above,

$$(a)_{[\alpha]} = a_{\bar\rho_1\bar\rho_2\bar\rho_3\rho_1^*\rho_2^*\rho_3^*} \neq a_{\bar\alpha\alpha^*},$$

in general. Also

$$(b_{\beta\gamma\gamma^*\beta^*})_{[\beta]} = b_{\rho_1\rho_2\bar\rho_1\bar\rho_2\rho_1^*\rho_2^*\bar\rho_1^*\bar\rho_2^*} = b_{\gamma\bar\beta\gamma^*\bar\beta^*} \neq b_{\bar\beta\gamma\gamma^*\bar\beta^*}$$

in general.

However, the rule that transposes of an element are obtained if the indices $\alpha_1, \ldots, \alpha_p$, or $\alpha_1^*, \ldots, \alpha_{p'}^*$, or $\beta_1, \ldots, \beta_q$, etc. are permuted among themselves, still holds.

A 1.3f Upper indices. A lower index possessing an asterisk may be replaced by an upper index which is the same but without the asterisk. Similarly, a lower index without an asterisk may be replaced by the corresponding upper index with an asterisk. Thus, if $a \in \mathcal{S}_{\alpha^3\alpha^{*2}\beta^2\beta^*\gamma}$ then we may write

$$a = a^{\alpha_1^*\ \alpha_3^*\alpha_1\ \ \beta_1^*\ \ \beta\gamma^*}_{\ \alpha_2\ \ \ \alpha_2^*\ \ \beta_2}$$

The indices would normally have to be staggered in this way to avoid

ambiguity. However, I shall make the convention of only writing upper indices if they are intended to be considered as occurring first and the lower ones afterwards. Hence, no staggering will be necessary. Thus if $b \in \mathcal{S}_{\alpha^1\alpha^*\beta^1\beta^*}$, then

$$b = b^{\alpha_1^*\beta_1^*\beta}{}_{\alpha_2\alpha^*\beta_2} = b^{\alpha_1^*\beta_1^*\beta}_{\alpha_2\,\alpha^*\beta_2} .$$

A 1.3g Contracted products. Let $a = a^{\alpha}_{\beta} \in \mathcal{S}_{\alpha^*\beta}$ and $b = b^{\beta}_{\gamma} \in \mathcal{S}_{\beta^*\gamma}$.

Then the contracted product a.b of a with b is given by

$$a.b = a^{\alpha}_{\beta}\ b^{\beta}_{\gamma} .$$

It will be seen that the definition of the contracted product of two elements of $\mathcal{S}$ depends on the particular divisions of indices into upper and lower indices which are chosen for the two elements. For this reason it will be necessary to associate the symbol a, not only with a particular element of the set $\mathcal{S}$, but also with a particular division of its indices into two groups, upper and lower. The same will hold for any other symbol representing an element of $\mathcal{S}$ (except that in the case of scalars this consideration is trivial.) If it is required to consider two different groupings of indices for the same element of $\mathcal{S}$, two different symbols must be used. This may be done by using the same letter in each case, but one of them being primed. Thus, if the element a, above, is required to be written with two lower indices, say, then we may write

$$a'_{\alpha^*\beta} = a^{\alpha}_{\beta}$$

A 1.3h It will be found convenient to make a slight alteration in the definition of the outer product of two elements of $\mathcal{S}$, when the symbols possess both upper and lower indices. Let a^{α}_{β}, $b^{\gamma}_{\delta} \in \mathcal{S}$ where

$\alpha = \rho^{t}\rho^{*t'}\sigma^{s} \ldots$, $\beta = \rho^{l}\rho^{*l'}\sigma^{m} \ldots$, $\gamma = \rho^{\wedge}\rho^{*\wedge'}\sigma^{a} \ldots$,

$\delta = \rho^{u}\rho^{*u'}\sigma^{v} \ldots$ and define $c^{\alpha\gamma}_{\beta\delta}$ by

$$c^{\rho_1 \cdots \rho_{\beta+\lambda}\, \rho^*_{\beta'+\lambda'}\, \sigma_1 \cdots \sigma_{\beta+\alpha} \cdots}_{\rho_{\beta+\lambda+1} \cdots \rho_{\beta'+\lambda'+l+u}\, \rho^*_{\beta'+\lambda'+1} \cdots \sigma_{\beta'+\alpha'+1} \cdots} = a^{\rho_1 \cdots \rho_{\beta}\, \rho^*_{\beta'}\, \sigma_1 \cdots \sigma_{\beta} \cdots}_{\rho_{\beta+\lambda+1} \cdots \rho_{\beta'+\lambda'+l}\, \rho^*_{\beta'+\lambda'+1} \cdots \rho^*_{\beta'+\lambda'+l'}\, \sigma_{\beta'+\alpha'+1} \cdots}\, b^{\rho_{\beta+1} \cdots \rho_{\beta+\lambda}\, \rho^*_{\beta'+1} \cdots \rho^*_{\beta'+\lambda'}\, \sigma_{\beta+1} \cdots \sigma_{\beta+\alpha} \cdots}_{\rho_{\beta'+\lambda'+1} \cdots \rho_{\beta'+\lambda'+l+u}\, \rho^*_{\beta'+\lambda'+l'+1} \cdots},$$

so that c is a transpose of the product given in A 1.2c and A 1.2m. The indices of c are arranged so that the upper indices of a occur first, followed by the upper indices of b, the lower indices of c then being the lower indices of a followed by the lower indices of b. This differs from the original definition of product (A 1.2c and A 1.2m) which would require that all the indices of a be taken first and those of b afterwards. The original definition can be regarded as giving rise to an expression of the form $d^{\alpha}{}_{\beta}{}^{\gamma}{}_{\delta}$, with staggered indices. This symbol would generally behave in the same way as $c^{\alpha\gamma}_{\beta\delta}$ with regard to sums and contracted products, but it does not satisfy the convention stated in A 1.3f, that the upper indices of d should be numbered before the lower ones. The definition of contracted product given in A 1.3g remains the same with this new definition of the outer product, since the indices which have to be interchanged are contracted with each other. This new definition

of product is clearly associative (A a.2c) and distributive over addition (A 1.2d) and satisfies laws corresponding to A 1.2o, A 1.2p and A 1.2t.

A 1.3i If a^{α}_{β} , b^{β}_{γ} , $c^{\gamma}_{\delta} \in \mathcal{S}$, then the associative law for contracted products

$$a\bullet(b\bullet c) = (a\bullet b)\bullet c$$

holds, as is clear from A 1.2c, A 1.2r and the definition A 1.3g. The brackets may, therefore, be omitted.

A 1.3j If a^{α}_{β} , b^{α}_{β} , c^{β}_{γ} , $d^{\delta}_{\alpha} \in \mathcal{S}$, then the distributive laws

$$(a + b)\bullet c = a\bullet c + b\bullet c, \quad d\bullet(a + b) = d\bullet a + d\bullet b$$

hold. This follows from A 1.2d and A 1.2u.

A 1.3k If a^{α}_{β} , b^{β}_{γ} , c^{δ}_{η} , $d^{\eta}_{\theta} \in \mathcal{S}$, then

$$(a\bullet b)(c\bullet d) = (ac)\bullet(bd) .$$

This follows from A 1.2p. The result is true only in virtue of the definition of multiplication given in A 1.3h.

A 1.3l Notation. In expressions involving both contracted products and outer products, the convention will be used that the outer product is the more binding with regard to the positioning of brackets. Thus, the right hand side of relation A 1.3k may be written $ac\bullet bd$.

A 1.3m The trace of an element. Let $a^{\alpha}_{\alpha} \in \mathcal{S}_{\alpha\alpha^*}$. Then the trace $\langle a \rangle$ of a is the scalar given by

$$\langle a \rangle \qquad = a \begin{matrix} \bar{\alpha} \\ \bar{\alpha} \end{matrix}$$

This definition, also, depends on a particular division of indices into upper and lower indices, since

$$\left\langle b \begin{matrix} \rho_1 \rho_1^* \\ \rho_2 \rho_2^* \end{matrix} \right\rangle = b \begin{matrix} \bar{\rho}_1 \bar{\rho}_1^* \\ \bar{\rho}_1 \bar{\rho}_2^* \end{matrix}$$

whereas if $b' \begin{matrix} \rho_1 \rho_2 \\ \rho_1 \rho_2 \end{matrix} = b \begin{matrix} \rho_1 \rho_1^* \\ \rho_2 \rho_2^* \end{matrix}$ we have

$$\langle b' \rangle = \left\langle b' \begin{matrix} \rho_1 \rho_2 \\ \rho_1 \rho_2 \end{matrix} \right\rangle = b' \begin{matrix} \bar{\rho}_1 \bar{\rho}_2 \\ \bar{\rho}_1 \bar{\rho}_2 \end{matrix} = b \begin{matrix} \bar{\rho}_1 \bar{\rho}_1^* \\ \bar{\rho}_2 \bar{\rho}_2^* \end{matrix}$$

which is generally distinct from $\langle b \rangle$.

A 1.3n If a_α^α, $b_\alpha^\alpha \in \mathfrak{J}$, then by A 1.2u,

$$\langle a + b \rangle = \langle a \rangle + \langle b \rangle$$

A 1.3o If a_α^α, $b_\beta^\beta \in \mathfrak{J}$, then by A 1.2p and A 1.2w,

$$\langle a \cdot b \rangle = \langle b \cdot a \rangle$$

(Hence, $\langle a.b.c \rangle = \langle b.c.a \rangle = \langle c.a.b \rangle$, etc.)

A 1.3p If a_α^α, $b_\beta^\beta \in \mathfrak{J}$, then by A 1.2t (and A 1.3h),

$$\langle ab \rangle = \langle a \rangle \langle b \rangle$$

A 1.3q If $a_\alpha^\alpha \in \mathfrak{J}$ and $\lambda \in \mathfrak{J}_1$, then

$$\langle \lambda a \rangle = \lambda \langle a \rangle .$$

This is a special case of A 1.3m.

A 1.3r As with the operation of multiplication (see A 1.3h), it will be convenient to redefine the meaning of the notation for contraction in the case of a symbol possessing both upper and lower indices. Thus if

$$a = a^{\rho_1 \cdots \tau_1 \cdots \tau_k \tau_1^* \cdots \tau_{k'}^* \cdots}_{\rho_1 \cdots \tau_{k'+1} \cdots \tau_{k'+k} \tau_{k+1}^* \cdots \tau_{k+k'}^* \cdots} \in \mathfrak{S}$$

then the symbol $(a)_{[\tau]}$ will be taken to mean

$$(a)_{[\tau]} = a^{\rho_1 \cdots \tau_1 \cdots \tau_{k-1} \overline{\tau} \tau_1^* \cdots \tau_{k'}^* \cdots}_{\rho_1 \cdots \tau_{k'+1} \cdots \tau_{k'+k-1} \overline{\tau} \tau_k^* \cdots \tau_{k+k'-1}^* \cdots}$$

rather than

$$a^{\rho_1 \cdots \tau_1 \cdots \tau_k \tau_1^* \cdots \tau_{k'}^* \cdots}_{\rho_1 \cdots \tau_{k+1} \cdots \tau_{k'+k-1} \overline{\tau} \tau_{k+1}^* \cdots \tau_{k+k'-1}^* \overline{\tau}^* \cdots}$$

That is, $(a)_{[\tau]}$ is obtained from a by contracting the highest numbered upper τ-index with the highest numbered lower τ-index of a. It should be noted that with this definition, in general

$$(a)_{[\tau^*]} \neq (a)_{[\tau]} ,$$

since a different pair of indices is contracted in each case.

The operation of α-contraction can also be defined in this way for an element with both upper and lower indices, as in A 1.3d. Thus, if $\alpha = \rho\rho \cdots \sigma \cdots \tau \cdots$, then $(a)_{[\alpha]} = \cdots(\cdots(\cdots((a)_{[\rho]})_{[\rho]} \cdots)_{[\sigma]} \cdots)_{[\tau]} \cdots$ etc. We have $((a)_{[\alpha]})_{[\beta]} = ((a)_{[\beta]})_{[\alpha]} = (a)_{[\alpha\beta]}$, $(a + b)_{[\alpha]} = (a)_{[\alpha]} + (b)_{[\alpha]}$, $a(b)_{[\alpha]} = (ab)_{[\alpha]}$, etc.

It will be noted that although, as in A 1.3d, the α-contraction of an element a is not always obtained by replacing the highest numbered upper index α and lower index α of a by dummy indices, the α-contraction is in fact so obtained, if the last appearing upper and lower indices of a are both α's. Thus, for example, if $a = a^{\alpha_2 \alpha_3 \gamma \alpha_1^* \alpha_1}_{\alpha_4^* \beta \alpha_2 \alpha_3}$, then

CHAPTER 6

$$(a)_{[\alpha]} = a^{\alpha_2 \bar{\alpha}\, \gamma \alpha_1^* \alpha_1}_{\alpha_4^* \beta \alpha_2 \bar{\alpha}}$$

This will hold, in particular, if all the indices of a are α's.

A 1.3e Notation. Let $a^{\alpha}_{\beta} \in \mathcal{S}$ and let r be a non-negative integer. Then the product aa....a of r a's in succession will be denoted by

$$(a)^r$$

the bracket being used to avoid possible confusion of r with an upper tensor index. If r = 0, the usual convention about zero exponents will be used, namely

$$(a)^0 = 1,$$

where 1 is the unit of $\mathcal{S}_1$.

A 1.4 The systems $\mathcal{J}$ and $\mathcal{R}$.

Two important subsystems of $\mathcal{S}$ will be described in this subsection. They will be of basic importance to section B ~~and C~~ and the rest of this chapter will be mainly devoted to a study of their properties.

A 1.4a Units of $\mathcal{J}$. Let $\alpha \in \mathfrak{S}$. Then the tensor system $\mathcal{S}$ may contain an element $I_{(\alpha)}{}^{\alpha}_{\alpha} \in \mathcal{S}_{\alpha\alpha^*}$, having the property that

$$a \bullet I_{(\alpha)} = a \quad \text{and} \quad I_{(\alpha)} \bullet b = b$$

for each $a^{\gamma}_{\alpha} \in \mathcal{S}_{\gamma^*\alpha}$, $b^{\alpha}_{\beta} \in \mathcal{S}_{\alpha^*\beta}$. Such an element will be called a unit of $\mathcal{S}$, or, more definitively, an α-unit of $\mathcal{S}$.

In the case of ordinary tensors, the units are Kronecker deltas.

A 1.4b If $\mathcal{S}$ contains an α-unit $I_{(\alpha)}$, this element is clearly unique, since $I'_{(\alpha)} \bullet I_{(\alpha)} = I'_{(\alpha)}$ implies $I'_{(\alpha)} = I_{(\alpha)}$

A 1.4c If $\mathcal{S}$ contains an α-unit, it must also contain an α^*-unit, this being a transpose of $I_{(\alpha)}$ with a different division of indices into upper and lower indices. Thus, if $\alpha = \rho^{\lambda} \rho^{*\lambda'} \sigma^{\alpha}_{\dots}$, the α^*-unit $I_{(\alpha^*)}$ is given by

$$I_{(\alpha^*)} {}^{\rho_1 \dots \rho_\lambda \; \rho_1^* \dots \rho_{\lambda'}^* \; \sigma_1 \dots \sigma_\alpha \; \dots}_{\rho_{\lambda+1} \dots \rho_{\lambda+\lambda'} \; \rho_{\lambda'+1}^* \dots \rho_{\lambda'+\lambda}^* \; \sigma_{\alpha+1} \dots \sigma_{\alpha+\alpha'} \; \dots} = I_{(\alpha)} {}^{\rho_{\lambda+1}^* \dots \rho_{\lambda+\lambda'}^* \; \rho_{\lambda'+1} \dots \rho_{\lambda'+\lambda} \; \sigma_{\alpha'+1}^* \dots \sigma_{\alpha'+\alpha}^* \dots}_{\rho_1^* \dots \rho_{\lambda'}^* \; \rho_1 \dots \rho_\lambda \; \sigma_1^* \dots \sigma_{\alpha'}^* \dots}$$

as is easily verified.

A 1.4d If $\mathcal{S}$ contains an α-unit and a β-unit, then

$$I_{(\alpha)} \; I_{(\beta)} = I_{(\alpha\beta)} \qquad \text{and} \qquad (I_{(\alpha)})^{\lambda} = I_{(\alpha^\lambda)}$$

The elements $(I_{(\alpha)})^{\lambda}$ will sometimes be written $I^{(\lambda)}_{(\alpha)}$. We have $I^{(0)}_{(\alpha)} = I_{(1)} = 1$, the unit of $\mathcal{S}_1$.

A 1.4e Dimension of elements of $\mathfrak{S}$. Let $\alpha \in \mathfrak{S}$ and suppose $\mathcal{S}$ contains an α-unit $I_{(\alpha)}$. The dimension of the index type α is then defined to be the scalar

$$\nu(\alpha) = \langle I_{(\alpha)} \rangle$$

A 1.4f If $\mathcal{S}$ contains an α-unit and a β-unit, then

$$\nu(\alpha)\, \nu(\beta) = \nu(\alpha\beta)$$

CHAPTER 6

This follows from A 1.4d and A 1.3o.

A 1.4g Definition. The tensor system $\mathfrak{R}$ is defined to be the smallest tensor subsystem of $\mathfrak{S}$ which contains all the units of $\mathfrak{S}$. It is thus the tensor system generated by the units of $\mathfrak{S}$. Similarly, the system $\mathfrak{F}$ is defined to be the tensor system generated by the units of $\mathfrak{S}$ and the elements of the field $\mathfrak{S}_1$.

The tensor operations in $\mathfrak{R}$ and $\mathfrak{F}$ are the same as those defined for $\mathfrak{S}$. The systems $\mathfrak{R}$ and $\mathfrak{F}$ are defined over the same semigroup $\mathfrak{G}$ as is $\mathfrak{S}$. The sets $\mathfrak{R}_\alpha$ and $\mathfrak{F}_\alpha$ are given, for each $\alpha \in \mathfrak{G}$, by

$$\mathfrak{R}_\alpha = \mathfrak{R} \cap \mathfrak{S}_\alpha , \quad \mathfrak{F}_\alpha = \mathfrak{F} \cap \mathfrak{S}_\alpha$$

A 1.4h None of the sets $\mathfrak{R}_\alpha$, $\mathfrak{F}_\alpha$ is vacuous since each contains a zero 0_α (see A 1.2b). In most cases, however, 0_α will be the only element of $\mathfrak{R}_\alpha$ and of $\mathfrak{F}_\alpha$ (and, in particular, when α is not of the form $\beta\beta^*$).

A 1.4i We have

$$\mathfrak{F}_1 = \mathfrak{S}_1 , \quad \mathfrak{R}_1 \subseteq \mathfrak{S}_1 .$$

The field $\mathfrak{R}_1$ of scalars of $\mathfrak{R}$ contains all the rational elements of $\mathfrak{S}_1$ (see A 1.2y) and all the dimensions $\nu(\alpha)$, $\nu(\beta)$, ... of the elements of $\mathfrak{G}$ which possess dimensions (see A 1.4e). The field $\mathfrak{R}_1$ will not generally be the whole of $\mathfrak{S}_1$. (A tensor subsystem of $\mathfrak{S}$ need not contain all the scalars of $\mathfrak{S}$.)

A 1.5 The system $\mathfrak{J}$

Suppose $\mathfrak{J}$ contains an α-unit, for some $\alpha \in \mathfrak{S}$. Then the subsystem $\mathfrak{J}$ contains all the elements of the form

$$A_{(\alpha)\,\alpha_1 \cdots \alpha_\lambda}^{\alpha_1 \cdots \alpha_\lambda} = \sum_k \lambda_{(k)}\, I_{(\alpha)\,\alpha_{k(1)}}^{\alpha_1}\, I_{(\alpha)\,\alpha_{k(2)}}^{\alpha_2} \cdots I_{(\alpha)\,\alpha_{k(\lambda)}}^{\alpha_\lambda} \qquad \text{A 1.5a}$$

where the summation extends over all permutations k of 1, .., r. If each $\lambda_{(k)} \in \mathfrak{R}_0$, then $A_{(\alpha)} \in \mathfrak{R}_0$.

A 1.5b If $\mathfrak{J}$ contains the β-unit $I_{(\beta)}$, then the expression A 1.5a defines another element $A_{(\beta)}$ of $\mathfrak{J}$ which is obtained simply by replacing all the α's which occur in the expression by β's. The symbol A will be considered as standing for the substitutional expression (symmetric group ring element) $\sum \lambda_{(k)} k$, determined by the right hand side of A 1.5a. Different elements of $\mathfrak{J}$ may then be represented by supplying A with different bracketed suffixes, e.g. $A_{(\alpha)}$, $A_{(\beta)}$, ... The element $A_{(\alpha)}$ will not always determine the substitutional expression A uniquely, however, because of identical relations satisfied by $I_{(\alpha)}$ Such identical relations will be discussed in A 1.7 and A 1.9.

A 1.5c Definition. The system $\mathfrak{Z}$ is defined as follows :

$\mathfrak{Z}$ is the union of disjoint sets $\mathfrak{Z}_{(0)}, \mathfrak{Z}_{(1)}, \mathfrak{Z}_{(2)}, \ldots$ where $\mathfrak{Z}_{(\lambda)}$ is isomorphic with the group ring of the symmetric groupsof order r! defined over the field $\mathfrak{F}_1$ (r = 0, 1, 2, ...). ($\mathfrak{Z}_{(0)}$ and $\mathfrak{Z}_{(1)}$ are

both isomorphic to $\mathfrak{J}_1$.) Elements of $\mathfrak{J}$ will be denoted by Roman capitals, A, B, C, ... ; the sum and product of two elements A, B belonging to the same set $\mathfrak{J}_{(\lambda)}$ will be denoted respectively by A + B and A • B. Each element of $\mathfrak{J}_{(\lambda)}$ is a linear combination over $\mathfrak{J}_1$ of basic permutations each of which may be denoted by a symbol such as

$$J_{[k(1), k(2), \ldots, k(\lambda)]},$$

k being a permutation of 1, .., r. The product of two permutations is given by

$$J_{[k(1), \ldots, k(\lambda)]} \cdot J_{[h(1), \ldots, h(\lambda)]} = J_{[h(k(1)), \ldots, h(k(\lambda))]} . \qquad \text{A 1.5d}$$

This extends by linearity to give the product of any two elements of $\mathfrak{J}_{(\lambda)}$ so that the distributive laws hold.

The system $\mathfrak{J}$ also admits a third operation, that of <u>outer product</u>. This will be denoted by juxtaposition of symbols. The outer product of two permutations is given by

$$J_{[k(1), \ldots, k(\lambda)]} J_{[h(1), \ldots, h(\sigma)]} = J_{[k(1), \ldots, k(\lambda), \lambda+h(1), \lambda+h(2), \ldots, \lambda+h(\sigma)]} \qquad \text{A 1.5e}$$

The outer product of an element of $\mathfrak{J}_{(\lambda)}$ with an element of $\mathfrak{J}_{(\sigma)}$ is thus an element of $\mathfrak{J}_{(\lambda+\sigma)}$. This is essentially the same operation as that defined in A 1.2n. The outer product of two elements of $\mathfrak{J}$ is then defined by linearity from A 1.5e so that the two distributive laws

$$A(B + C) = AB + AC, \quad (B + C)A = BA + CA \qquad \text{A 1.5f}$$

hold for all B, C $\in \mathfrak{J}_{(\lambda)}$, A $\in \mathfrak{J}_{(\lambda)}$. With this definition, it follows that the two laws

$$B(CD) = (BC)D, \quad (A \bullet E)(B \bullet C) = (AB) \bullet (EC) \qquad \text{A 1.5g}$$

hold for all B, C, D $\in \mathfrak{J}_{(\lambda)}$ and A, E $\in \mathfrak{J}_{(\epsilon)}$, since they hold for permutations with the definition A 1.5e. Using the convention A 1.31, the brackets may be omitted from the right hand side of the second relation in A 1.5g. (They may, of course, also be omitted in the product BCD, etc.)

A 1.5h The systems $\mathfrak{F}^{(\alpha)}$ and $\mathfrak{R}^{(\alpha)}$. For each $\alpha \in \mathfrak{S}$ for which $\mathfrak{J}$ contains an α-unit, it is possible to set up a mapping $\Theta^{(\alpha)}$ of $\mathfrak{J}$ into $\mathfrak{F}$. The element $J_{[k(1), \cdots, k(\lambda)]}$ of $\mathfrak{J}$ is mapped on to the element

$$J_{[k(1), \cdots, k(\lambda)](\alpha)} = I_{(\alpha)\alpha_{k(1)}}^{\alpha_1} \; I_{(\alpha)\alpha_{k(2)}}^{\alpha_2} \cdots I_{(\alpha)\alpha_{k(\lambda)}}^{\alpha_\lambda}$$

The mapping of the other elements of $\mathfrak{J}$ is then defined by linearity over $\mathfrak{J}_1$ and gives rise to elements such as $A_{(\alpha)}$ defined in A 1.5a. The system $\mathfrak{F}^{(\alpha)}$ is defined to be the image of $\mathfrak{J}$ under $\Theta^{(\alpha)}$. The system $\mathfrak{R}^{(\alpha)}$ is defined by

$$\mathfrak{R}^{(\alpha)} = \mathfrak{F}^{(\alpha)} \cap \mathfrak{R}$$

The mapping $\Theta^{(\alpha)}$ is clearly a homomorphism with respect to addition. It is also a homomorphism with respect to the two operations of multiplication, as will be seen from the definitions A 1.5d and A 1.5e.

$$(J_{[k(1),\ldots,k(\lambda)](\alpha)} \cdot J_{[h(1),\ldots,h(\lambda)](\alpha)} = I_{(\alpha)\bar{\alpha}_{k(1)}}^{\alpha_1} \ldots I_{(\alpha)\bar{\alpha}_{k(\lambda)}}^{\alpha_\lambda} I_{(\alpha)\alpha_{h(1)}}^{\bar{\alpha}_1} \ldots I_{(\alpha)\alpha_{h(\lambda)}}^{\bar{\alpha}_\lambda}$$

$$= I_{(\alpha)\alpha_{h(k(1))}}^{\alpha_1} \ldots I_{(\alpha)\alpha_{h(k(\lambda))}}^{\alpha_\lambda} = J_{[h(k(1)),\ldots,h(k(\lambda))]}.)$$

The sets $\mathcal{J}_1^{(\alpha)}$ and $\mathcal{R}_1^{(\alpha)}$ are, respectively, the field $\mathcal{J}_1$ and the field $\mathcal{R}_1$.

A 1.5i The systems $\mathcal{J}^{(\alpha)}$ and $\mathcal{R}^{(\alpha)}$ are not tensor systems since they only contain elements in those sets $\mathcal{J}_\beta$ for which β is of the form $\alpha \wedge \alpha^* \sim$ (see A 1.4h). However, if α is a generator of $\mathfrak{S}$, the set $\mathcal{J}^{(\alpha)}$ (or $\mathcal{R}^{(\alpha)}$) together with the zeros of $\mathcal{J}$ does in fact constitute a tensor system. This follows because any transpose or contraction of an element $J_{[k(1),\ldots](\alpha)}$ is again such an element, possibly multiplied by $\nu(\alpha)$ (which is an element of $\mathcal{R}_1^{(\alpha)}$).

A 1.5j If $\mathcal{J}$ contains an α-unit for each $\alpha \in \mathfrak{S}$, then any transpose or contraction of an element a_β of $\mathcal{J}$ can be written in the form $a \cdot F$ where F is some element of $\mathcal{R}$. For let $a_\beta = a_{\rho_1 \ldots \rho_\lambda \rho_1^* \ldots \rho_{\lambda'}^* \sigma_1 \ldots \sigma_\alpha \ldots}$; then

$$a_{\rho_{k(1)} \ldots \rho_{k(\lambda)} \rho_{k'(1)}^* \ldots \rho_{k'(\lambda')}^* \sigma_{\ell(1)} \ldots \sigma_{\ell(\alpha)} \ldots} = a \cdot J_{[k(1),\ldots,k(\lambda)](\rho)} J_{[k'(1),\ldots,k'(\lambda')](\rho^*)} J_{[\ell(1),\ldots,\ell(\alpha)](\sigma)} \ldots$$

and

$$a_{\rho_1 \ldots \rho_{\lambda-1} \rho_1^* \ldots \rho_{\lambda'-1}^* \bar{\rho} \bar{\rho}^* \sigma_1 \ldots \sigma_\alpha \ldots} = a \cdot I_{(\rho)}^{(\lambda-1)} I_{(\rho^*)}^{(\lambda'-1)} I_{(\rho)}^{\prime} I_{(\sigma)}^{(\alpha)} I_{(\sigma^*)}^{(\alpha')}$$

where

$$I_{(\rho)}^{\prime\, \rho \rho^*} = I_{(\rho)}^{\rho}$$

etc.

A 1.5k Notation. For convenience, I shall frequently omit the suffix (α) of a symbol $I_{(\alpha)}$ or $A_{(\alpha)}$, etc., but generally only if it is the particular Greek letter α that is involved (or otherwise if there is no possible ambiguity ; e.g. if $e = e^{\beta}_{\beta}$, then $I - e$ means $I_{(\beta)} - e$).

A 1.5l Definition. Let

$$P^{(\lambda)} = \frac{1}{\lambda!} \sum_{k} J_{[k(1),\ldots,k(\lambda)]}$$

and

$$N^{(\lambda)} = \frac{1}{\lambda!} \sum_{k} \eta_k J_{[k(1),\ldots,k(\lambda)]}$$

where the sums are taken over all permutations k of 1, .., r and where η_k is defined to be +1 or -1 according as k is an even or an odd permutation.

(These are Young's operators "the positive symmetric group on r symbols" and "the negative symmetric group on r symbols" respectively, see Rutherford (21) p.p. 8,9.)

We have $P^{(0)} = N^{(0)} = 1$ and $P^{(1)} = N^{(1)} = I$.

These definitions may be written formally as

$$P^{(\lambda)}_{(\alpha)} = \frac{1}{\lambda!} \begin{vmatrix} I^{\alpha_1}_{\alpha_1} & \cdots & I^{\alpha_\lambda}_{\alpha_1} \\ \vdots & & \vdots \\ I^{\alpha_1}_{\alpha_\lambda} & \cdots & I^{\alpha_\lambda}_{\alpha_\lambda} \end{vmatrix}^{+} \quad \text{and} \quad N^{(\lambda)}_{(\alpha)} = \frac{1}{\lambda!} \begin{vmatrix} I^{\alpha_1}_{\alpha_1} & \cdots & I^{\alpha_\lambda}_{\alpha_1} \\ \vdots & & \vdots \\ I^{\alpha_1}_{\alpha_\lambda} & \cdots & I^{\alpha_\lambda}_{\alpha_\lambda} \end{vmatrix} . \qquad \text{A 1.5m}$$

In the case of ordinary tensors, the symbol $r!N^{(r)}_{(\alpha)}$ represents the generalized Kronecker delta $\delta^{i_1 \cdots i_\lambda}_{j_1 \cdots j_\lambda}$ (see Hodge (7) p.21 and also Eisenhart (5) p.107)

A 1.5n Notation. If the symbols $P^{(\lambda)}_{(\beta)}$ and $N^{(\lambda)}_{(\beta)}$ are written with their upper and lower indices $\beta_1 \cdots \beta_\lambda$, it will be convenient to omit the (r) of the symbol as well as the (β). Thus,

$$P^{\beta_1 \cdots \beta_\lambda}_{\beta_1 \cdots \beta_\lambda} = P^{(\lambda)}_{(\beta)} \qquad \text{and} \quad N^{\beta_1 \cdots \beta_\lambda}_{\beta_1 \cdots \beta_\lambda} = N^{(\lambda)}_{(\beta)} \ .$$

A 1.5o Definition. Let $a^{\beta}_{\alpha^\lambda} \in \mathcal{J}$; then the element a will be called α-symmetric in its lower indices if the transpose obtained by interchanging any two of the indices $\alpha_1, \ldots, \alpha_\lambda$ of the symbol $a^{\beta}_{\alpha_1 \cdots \alpha_\lambda}$ is again the element a. It will be called α-skew-symmetric in its lower indices if such an interchange always produces the element-a. Similarly, an element $b^{\alpha^\lambda}_{\beta}$ of $\mathcal{J}$ may be α-symmetric or α-skew-symmetric in its upper indices.

A 1.5p It will be seen that the elements $P^{(\lambda)}_{(\alpha)}$ are α-symmetric in their lower indices and also in their upper indides. For, interchanging two lower indices or two upper indices merely interchanges pairs of terms in the sum given in A 1.5l. Similarly, the elements $N^{(\lambda)}_{(\alpha)}$ are α-skew-symmetric in their lower and in their upper indices.

A 1.5q Lemma. If $a^{\beta}_{\alpha^\lambda}$, $b^{\alpha^\lambda}_{\beta} \in \mathcal{J}$, then necessary and sufficient conditions for a to be α-symmetric or α-skew-symmetric in its lower

lower indices are, respectively

$$a \bullet P^{(\lambda)}_{(\alpha)} = a, \quad a \bullet N^{(\lambda)}_{(\alpha)} = a.$$

The corresponding conditions for the upper indices of b are, respectively

$$P^{(\lambda)}_{(\alpha)} \bullet b = b, \quad N^{(\lambda)}_{(\alpha)} \bullet b = b.$$

Proof. If $a^{\beta}_{\alpha_1 \cdots \alpha_\lambda}$ is α-symmetric in its lower indices then any permutation of $\alpha_1, \cdots, \alpha_\lambda$ leaves a unchanged. Hence, for each permutation k,

$$a \bullet J_{[k(1), \cdots, k(\lambda)]} = a,$$

so that $a \bullet P^{(\lambda)} = \frac{1}{\lambda!} \sum_k a \bullet J_{[k(1), \cdots, k(\lambda)]} = a$. Conversely, if $a = a \bullet P^{(\lambda)}$, then a must be α-symmetric in its lower indices since $P^{(\lambda)}$ is, by A 1.5p. The proof in the case of α-skew-symmetric indices is similar. The result for upper indices is essentially the same as that for lower indices.

Applying this proof to the elements $P^{(\delta)}$ and $N^{(\delta)}$, results such as

$$\left.\begin{aligned} P^{(\delta)} \bullet I^{(t)} P^{(\lambda)} I^{(\delta - t - \lambda)} = I^{(t)} P^{(\lambda)} I^{(\delta - t - \lambda)} \bullet P^{(\delta)} = P^{(\delta)} = P^{(\delta)} \bullet P^{(\delta)} \\ N^{(\delta)} \bullet I^{(t)} N^{(\lambda)} I^{(\delta - t - \lambda)} = I^{(t)} N^{(\lambda)} I^{(\delta - t - \lambda)} \bullet N^{(\delta)} = N^{(\delta)} = N^{(\delta)} \bullet N^{(\delta)} \end{aligned}\right\} \text{A 1.5r}$$

may be obtained ($s \geqslant r + t$).

A 1.5s Lemma. If $r > 0$, then

$$P^{(\lambda)}_{(\alpha)} = \frac{1}{\lambda} \sum_{i=1}^{\lambda} P^{\alpha_1 \cdots \alpha_{i-1} \alpha_{i+1} \cdots \alpha_\lambda}_{\alpha_1 \alpha_2 \cdots \alpha_{\lambda-1}} I^{\alpha_i}_{\alpha_\lambda} \quad \text{and} \quad N^{(\lambda)}_{(\alpha)} = \frac{1}{\lambda} \sum_{i=1}^{\lambda} (-1)^{\lambda - i} N^{\alpha_1 \cdots \alpha_{i-1} \alpha_{i+1} \cdots \alpha_\lambda}_{\alpha_1 \alpha_2 \cdots \alpha_{\lambda-1}} I^{\alpha_i}_{\alpha_\lambda}$$

Proof. Consider the case of $N^{(\lambda)}_{(\alpha)}$. Now, by A 1.5e, we have

$$N^{\alpha_1 \cdots \alpha_{i-1} \alpha_{i+1} \cdots \alpha_\lambda}_{\alpha_1 \alpha_2 \cdots \alpha_{\lambda-1}} I^{\alpha_i}_{\alpha_\lambda} = \frac{1}{(\lambda-1)!} \sum_k \eta_k I^{\alpha}_{\alpha_{k(1)}} \cdots I^{\alpha_{i-1}}_{\alpha_{k(i-1)}} I^{\alpha_i}_{\alpha_\lambda} I^{\alpha_{i+1}}_{\alpha_{k(i)}} \cdots I^{\alpha_\lambda}_{\alpha_{k(\lambda-1)}}$$

$$= \frac{1}{(\lambda-1)!} \sum_k \eta_k J_{[k(1), \ldots, k(i-1), \alpha, k(i), \ldots, k(\lambda-1)]}$$

where k represents a permutation of 1,.., r-1 and η_k is +1 or -1 according as k is even or odd. But any permutation of 1, .., r is of the form k(1),..,k(i-1) ,r, k(i),..,k(λ-1) and is even or odd according as $(-1)^{\lambda-k}$ η_k is +1 or -1. The result follows. (See also Rutherford (20) p.10.)

This result may also be obtained from a formal Laplacian expansion of the determinant of A 1.5m, with respect to the last row.

The proof for $P^{(\lambda)}_{(\alpha)}$ is similar.

A 1.5t Lemma. If r > 0, then

$$P^{(\lambda)}_{(\alpha)} = \frac{1}{\lambda} P^{(\lambda-1)}_{(\alpha)} I_{(\alpha)} + \frac{\lambda-1}{\lambda} P^{\alpha_1 \cdots \alpha_{\lambda-1}}_{\bar{\alpha}_1 \cdots \bar{\alpha}_{\lambda-2} \alpha_\lambda} P^{\bar{\alpha}_1 \cdots \bar{\alpha}_{\lambda-2} \alpha_\lambda}_{\alpha_1 \cdots \alpha_{\lambda-1}} \quad \text{and}$$

$$N^{(\lambda)}_{(\alpha)} = \frac{1}{\lambda} N^{(\lambda-1)}_{(\alpha)} I_{(\alpha)} - \frac{\lambda-1}{\lambda} N^{\alpha_1 \cdots \alpha_{\lambda-1}}_{\bar{\alpha}_1 \cdots \bar{\alpha}_{\lambda-2} \alpha_\lambda} N^{\bar{\alpha}_1 \cdots \bar{\alpha}_{\lambda-2} \alpha_\lambda}_{\alpha_1 \cdots \alpha_{\lambda-1}}$$

Proof. Consider the case of $N^{(\lambda)}_{(\alpha)}$. By A 1.5r (and A 1.3k) we have

$N^{(\lambda)} = N^{(\lambda-1)} I \bullet N^{(\lambda)}$ and $N^{(\lambda-1)} I = N^{(\lambda-1)} I \bullet N^{(\lambda-1)} I$. Hence, by A 1.5s,

$$N^{(\lambda)}_{(\alpha)} = \frac{1}{\lambda} N^{(\lambda-1)}_{(\alpha)} I_{(\alpha)} + \frac{1}{\lambda} \sum_{i=1}^{\lambda-1} (-1)^{\lambda-i} N^{\alpha_1 \cdots \alpha_{\lambda-1}}_{\bar{\alpha}_1 \cdots \bar{\alpha}_{\lambda-1}} N^{\bar{\alpha}_1 \cdots \bar{\alpha}_{i-1} \bar{\alpha}_{i+1} \cdots \bar{\alpha}_{\lambda-1} \alpha_\lambda}_{\alpha_1 \alpha_2 \cdots \alpha_{\lambda-1}} I^{\bar{\alpha}_i}_{\alpha_\lambda}$$

and the result follows by the skew symmetry of $N^{(\lambda-1)}$. The proof for $P^{(\lambda)}_{(\alpha)}$ is similar.

A 1.5u Lemma. Let $\nu = \nu(\alpha)$ and r ⩾ s ⩾ 0. Then

$$\left(P^{(\lambda)}_{(\alpha)}\right)_{[\alpha^{\lambda-\delta}]} = \frac{(\nu+\lambda-1)(\nu+\lambda-2)\cdots(\nu+\delta)}{\lambda(\lambda-1)\cdots(\delta+1)} P^{(\delta)}_{(\alpha)} \qquad \text{and}$$

$$\left(N^{(\lambda)}_{(\alpha)}\right)_{[\alpha^{\lambda-\delta}]} = \frac{(\nu-\lambda+1)(\nu-\lambda+2)\cdots(\nu-\delta)}{\lambda(\lambda-1)\cdots(\delta+1)} N^{(\delta)}_{(\alpha)} .$$

<u>Proof</u> From A 1.5t we get $P^{\alpha_1\cdots\alpha_{\lambda-1}\overline{\alpha}}_{\alpha_1\cdots\alpha_{\lambda-1}\overline{\alpha}} = \frac{\nu}{\lambda} P^{(\lambda-1)} + \frac{\lambda-1}{\lambda} P^{(\lambda)} \cdot P^{(\lambda)}$

$$= \frac{\nu+\lambda-1}{\lambda} \cdot P^{(\lambda)}$$

by A 1.5r. Similarly, $N^{\alpha_1\cdots\alpha_{\lambda-1}\overline{\alpha}}_{\alpha_1\cdots\alpha_{\lambda-1}\overline{\alpha}} = \frac{\nu-\lambda+1}{\lambda} N^{(\lambda)}$. The result follows.

A 1.6 <u>Normal tensor systems</u>

A tensor system $\mathcal{S}$ will be called <u>normal</u> if it possesses an operation, denoted by $*$, on its elements, with the following properties :

A 1.6a If $a = a^{\rho_1\cdots\rho_\lambda\,\sigma_1\cdots\sigma_\delta\,\cdots}_{\rho_1\cdots\rho_\lambda\,\sigma_1\cdots\sigma_\delta\,\cdots} \in \mathcal{S}_\alpha$, then

$a^* = a^{*\,\rho_1\cdots\rho_\lambda\,\sigma_1\cdots\sigma_\delta\,\cdots}_{\rho_1\cdots\rho_\lambda\,\sigma_1\cdots\sigma_\delta\,\cdots} \in \mathcal{S}_{\alpha^*}$ ($\rho, \sigma, \ldots$ being a set of semigenerators of $\mathfrak{S}$),

A 1.6b $(a^*)^* = a$,

A 1.6c $(a + b)^* = a^* + b^*$,

A 1.6d $(ab)^* = a^* b^*$,

A 1.6e $(a_{[\tau]})^* = (a^*)_{[\tau]}$,

A 1.6f $(a^{\rho_{k(1)}\cdots\rho_{k(\lambda)}\,\sigma_{\ell(1)}\cdots\sigma_{\ell(\delta)}\,\cdots}_{\rho_{k'(1)}\cdots\rho_{k'(\lambda')}\,\sigma_{\ell'(1)}\cdots\sigma_{\ell'(\delta')}\,\cdots})^* = a^{*\,\rho_{k(1)}\cdots\rho_{k(\lambda)}\,\sigma_{\ell(1)}\cdots\sigma_{\ell(\delta)}\,\cdots}_{\rho_{k'(1)}\cdots\rho_{k'(\lambda')}\,\sigma_{\ell'(1)}\cdots\sigma_{\ell'(\delta')}\,\cdots}$

for each permutation of k of 1, .., r ; k' of 1, .., r' ; l of 1, .., s ; etc.

A 1.6g If $a \neq 0$, then $a^{\bar\rho_1 \cdots \bar\rho_{\kappa'}\ \bar\sigma_1 \cdots \bar\sigma_{\alpha'}\ \cdots}_{\bar\rho_{\kappa'+1} \cdots \bar\rho_{\kappa'+\lambda}\ \bar\sigma_{\alpha'+1} \cdots \bar\sigma_{\alpha'+\alpha}\ \cdots}\ a^{*\ \bar\rho_{\kappa'+1} \cdots \bar\rho_{\kappa'+\lambda}\ \bar\sigma_{\alpha'+1} \cdots \bar\sigma_{\alpha'+\alpha}\ \cdots}_{\bar\rho_1 \cdots \bar\rho_{\kappa'}\ \bar\sigma_1 \cdots \bar\sigma_{\alpha'}\ \cdots} \neq 0.$

The element a^* will be called the <u>conjugate</u> of a. The operation $a \to a^*$ will be called an <u>involution</u> of $\mathfrak{J}$.

A 1.6h As was the case with multiplication (see A 1.3h) it will be convenient to make a slight alteration of the definition of a^* in the case of a symbol with a general division into upper and lower indices, in accordance with the convention of A 1.3f.

Let

$$a^{\rho_1 \cdots \rho_\kappa\ \rho_1^* \cdots \rho_\kappa^*\ \sigma_1 \cdots \sigma_\alpha\ \cdots}_{\rho_{\kappa+1} \cdots \rho_{\kappa'+\mu}\ \rho_{\kappa'+1}^* \cdots \rho_{\kappa'+\mu}^*\ \sigma_{\alpha'+1} \cdots \sigma_{\alpha'+\nu}\ \cdots} = b^{\rho_1 \cdots \rho_{\kappa+\mu'}\ \sigma_1 \cdots \sigma_{\alpha+\nu'}\ \cdots}_{\rho_1 \cdots \rho_{\kappa'+\mu}\ \sigma_1 \cdots \sigma_{\alpha'+\nu}\ \cdots}$$

then the conjugate a^* of a is that transpose of b^* given by

$$a^* = a^{*\ \rho_1 \cdots \rho_\mu\ \rho_1^* \cdots \rho_{\mu'}^*\ \sigma_1 \cdots \sigma_\nu\ \cdots}_{\rho_{\mu'+1} \cdots \rho_{\mu'+\lambda}\ \rho_{\mu+1}^* \cdots \rho_{\mu+\lambda'}^*\ \sigma_{\nu+1} \cdots \sigma_{\nu+\alpha'}\ \cdots} = b^{*\ \rho_{\mu+1} \cdots \rho_{\mu+\lambda'}\ \rho_1 \cdots \rho_\mu\ \sigma_{\nu+1} \cdots \sigma_{\nu+\alpha'}\ \sigma_1 \cdots \sigma_\nu\ \cdots}_{\rho_{\mu'+1} \cdots \rho_{\mu'+\lambda}\ \rho_1 \cdots \rho_{\mu'}\ \sigma_{\nu'+1} \cdots \sigma_{\nu'+\alpha}\ \sigma_1 \cdots \sigma_{\nu'}\ \cdots}.$$

When $r' = 0$, $u' = 0$, $s' = 0$,... this reduces to the original definition A 1.6a. With this extended definition, the upper indices of a^* all arise from the lower indices of a and the lower indices of a^* all arise from the upper indices of a and, in accordance with A 1.3f, the upper indices of a^* are all numbered first.

It may be verified that the laws A 1.6b,..,A 1.6e all remain true (using the definitions of multiplication and contraction given in A 1.3h ~~1.3h~~ and A 1.3r). A 1.6f becomes

A 1.6i $$\left(a^{\rho_{f(1)} \cdots \rho_{f(\kappa)}\ \rho_{\kappa+1}^* \cdots \rho_{\kappa+\lambda}^*\ \sigma_{g(1)} \cdots}_{\rho_{\kappa+f(1)} \cdots \rho_{\kappa+f(\kappa)}\ \rho_{\kappa+f(\kappa)+1}^* \cdots \rho_{\kappa+f(\kappa)+\mu}^*\ \sigma_{\alpha+g(1)} \cdots}\right)^* = a^{*\ \rho_{f(1)} \cdots \rho_{f(\kappa)}\ \rho_{f(\kappa)+1}^* \cdots \rho_{f(\kappa)+\mu}^*\ \sigma_{g(1)} \cdots}_{\rho_{\alpha+k(1)} \cdots \rho_{\alpha+k(\lambda)}\ \rho_{\alpha+k(\lambda)+1} \cdots \rho_{\alpha+k(\lambda)+\nu}^*\ \sigma_{\beta+l(1)} \cdots}$$

Also, A 1.6g may be written :

A 1.6j If $a^{\alpha}_{\beta} \neq 0^{\alpha}_{\beta}$, then $a^{\bar{\alpha}}_{\bar{\beta}}\, a^{*\bar{\beta}}_{\bar{\alpha}} \neq 0$;

i.e. $\langle a \cdot a^* \rangle = 0$ implies $a = 0$.

From A 1.6j (and A 1.2v) we deduce

A 1.6k $a \cdot a^* = 0$ implies $a = 0$.

Similarly, $a^* \cdot a = 0$ implies $a = 0$.

From A 1.6d, A 1.6e and A 1.2p (and A 1.6h) we get

A 1.6l $(a \cdot b)^* = b^* \cdot a^*$.

A 1.6m $a^* \cdot a \cdot b = a^* \cdot a \cdot c$ implies $a \cdot b = a \cdot c$.

This follows from A 1.6k since $(b - c)^* \cdot (a^* \cdot a \cdot b - a^* \cdot a \cdot c)$

$= (a \cdot b - a \cdot c)^* \cdot (a \cdot b - a \cdot c)$

Supposing $\mathfrak{S}$ to be normal, we have the following properties :

A 1.6n $(\mathfrak{S}_1)^* = \mathfrak{S}_1$ by A 1.6a. Thus, by A 1.6b, A 1.6c and A 1.6d, the mapping $\lambda \to \lambda^*$ ($\lambda \in \mathfrak{S}_1$) is an involution, or involutory (anti-)isomorphism, of the field $\mathfrak{S}$. The elements of the rational subfield of $\mathfrak{S}_1$ are all self-conjugate.

A 1.6o If $\mathfrak{S}$ possesses an α -unit, then $(I_{(\alpha)})^* = I_{(\alpha)}$

This follows from A 1.4b.

A 1.6p We have $(\mathfrak{R})^* = \mathfrak{R}$ and $(\mathfrak{F})^* = \mathfrak{F}$; that is, $\mathfrak{R}$ and $\mathfrak{F}$ are both normal with the same involution as $\mathfrak{S}$.

This follows from A 1.6o and A 1.6n.

A 1.6q If $\mathfrak{J}$ possesses an α-unit and k and k^{-1} are permutations of $1,\ldots,r$, k^{-1} being inverse to k, then

$$(J_{[k(1),\ldots,k(\omega)]})^* = J_{[k^{-1}(1),\ldots,k^{-1}(\omega)]}$$

(J is defined in A 1.5h).

This follows from A 1.6e, A 1.6d and A 1.6i.

A 1.6r If $\mathfrak{J}$ possesses an α-unit, we have

$$(P^{(\lambda)}_{(\alpha)})^* = P^{(\lambda)}_{(\omega)} \quad \text{and} \quad (N^{(\lambda)}_{(\alpha)})^* = N^{(\lambda)}_{(\alpha)}$$

as follows from A 1.5l.

A 1.6s Definition. Let $\mathfrak{J}_1$ be the field of complex numbers or a subfield of the complex field. Then if $\mathfrak{J}$ is normal and its involution $a \to a^*$ sends each element λ of $\mathfrak{J}_1$ into its complex conjugate $\bar{\lambda}$, $\mathfrak{J}$ will be called complex-normal.

It will be seen that the system of all the ordinary tensors with complex elements over a space, in fact, constitutes a complex-normal system. The conjugate of a tensor is obtained by interchanging covariant and contravariant indices and taking the complex conjugates of the elements. This operation is an involution satisfying A 1.6a,.., A 1.6g.

A 1.6t Definition. A tensor system $\mathfrak{J}$ (not necessarily normal) will be called weakly regular if for each $a_\alpha \neq 0$ of $\mathfrak{J}_\alpha$ there exists an element b^α of $\mathfrak{J}_{\alpha^*}$ such that

$$a_{\alpha} b^{\overline{\alpha}} \neq 0.$$

Clearly any normal tensor system is weakly regular since we may take $b = a^*$ (by A 1.6g). Regular systems will be defined in A 2.1.

A 1.7 **Theorem.** Suppose $\mathcal{S}$ possesses an α-unit. If, for some $n > 0$,

$$N^{(n+1)}_{(\alpha)} = 0 \quad \text{and} \quad N^{(n)}_{(\alpha)} \neq 0,$$

then the dimension of α (see A 1.4e) is given by

$$\nu(\alpha) = n.$$

Conversely, if $\mathcal{R}$ is normal and $\nu(\alpha)$ is the positive integer n, then $N^{(n+1)}_{(\alpha)} = 0$ and $N^{(n)}_{(\alpha)} \neq 0$.

Proof. Let $N^{(n+1)}_{(\alpha)} = 0$, $N^{(n)} \neq 0$. Now by A 1.5u we have

$$(N^{(n+1)})_{[\alpha]} = \frac{\nu(\alpha) - n}{n+1} N^{(n)},$$

whence, by A 1.2f, we have $\nu(\alpha) = n$. Conversely, suppose that $\mathcal{R}$ is normal and that $\nu(\alpha) = n$. Then we have $(N^{(n+1)})_{[\alpha]} = 0$. But, by A 1.5 and A 1.6r, $N^{(n+1)} = N^{(n+1)} \cdot (N^{(n+1)})^*$. Hence,

$$(N^{(n+1)} \cdot (N^{(n+1)})^*)_{[\alpha]} = 0.$$

This implies

$$\langle N^{(n+1)} \cdot (N^{(n+1)})^* \rangle = 0,$$

whence $N^{(n+1)} = 0$, by A 1.6j. Also $N^{(n)} \neq 0$, since, by A 1.5u,

$$\langle N^{(n)} \rangle = \frac{1.2 \ldots n}{n(n-1) \ldots 2.1} . 1 = 1 \neq 0.$$

A 1.7a Proper tensor systems. Let $\mathcal{S}$ be a tensor system over $\mathfrak{S}$ which possesses an α-unit for each $\alpha \in \mathfrak{S}$. Suppose that the subsystem $\mathcal{R}$ of $\mathcal{S}$ is normal and that the dimension $\nu(\alpha)$ of each $\alpha \in \mathfrak{S}$ is a positive integer. The tensor system $\mathcal{S}$ will then be called proper. (By A 1.4c, A 1.4d and A 1.4f, a sufficient characterization of a proper system is that, for a set $\rho, \sigma, \ldots$ of semigenerators of $\mathfrak{S}$, the units $I_{(\rho)}$, $I_{(\sigma)}$, .. exist, $\mathcal{R}$ is normal and the dimensions $\nu(\rho), \nu(\sigma), \ldots$ are positive integers.)

A 1.7b It is clear that proper tensor systems exist. For the system of all ordinary tensors over a given complex vector space is a complex normal system (see A 1.6s) so that $\mathcal{R}$ is normal (by A 1.6p). It possesses a unit for each index type (Kronecker deltas) and the dimensions of the various index types are the various powers of the dimension of the original vector space and are thus positive integers.

Similarly, the system of all ordinary tensors with elements taken from a field $\mathcal{R}$, without characteristic, over given space defined over $\mathcal{R}$, is also a proper tensor system. Its subsystem $\mathcal{R}$ is isomorphic with that of the system considered above, if the spaces have equal dimension, since the elements of each tensor belonging to $\mathcal{R}$ must be the same rational numbers in the two cases. (They are constructable by means of rational operations from the components 0, 1 of the Kronecker deltas.)

A 1.7c Let $\mathcal{S}$ be a proper tensor system. Then by theorem A 1.7c we see that the units satisfy certain identities (given by $N_{(\alpha)}^{(\nu(\alpha)+1)} = 0$). Hence, the mapping $\Theta^{(\alpha)}$ of $\mathcal{S}$ into $\mathcal{S}^{(\alpha)}$ (see A 1.5h) is not an isomorphism. Elements of $\mathcal{R}$ (or of $\mathcal{S}$) may be written down which do not vanish formally but which are zero nevertheless. However, if the integers $\nu(\rho)$, $\nu(\sigma)$, .. are known for the semigenerators of $\mathfrak{S}$, an algorithm can be given for deciding whether or not an element of $\mathcal{R}$ is a zero.

Suppose $A \in \mathcal{R}$ is given by

$$A = \sum_{k,l,..} \lambda_{k,l,..} I^{\rho_1}_{\rho_{k(1)}} .. I^{\rho_\lambda}_{\rho_{k(\lambda)}} I^{\sigma_1}_{\sigma_{l(1)}} .. I^{\sigma_\mu}_{\sigma_{l(\mu)}} ....$$

where k, l, .. are permutations and where the $\lambda_{k,l,..}$ are rational numbers. Any element of $\mathcal{R}$ must be of this form, since the set of elements of this form contains the units of $\mathcal{S}$ and it constitutes a tensor system, as is easily verified. (Contraction merely introduces the scalars $\nu(\rho)$, $\nu(\sigma)$, ,. and these are integers.) Now, by A 1.6g (or A 1.6j), the element A is zero if and only if $\langle A \cdot A^* \rangle = 0$. But

$$\langle A \cdot A^* \rangle = \sum_{k,l,..} \sum_{k',l',..} \lambda_{k,l,..} \lambda_{k',l',..} I^{\bar{\rho}_{k'(1)}}_{\bar{\rho}_{k(1)}} .. I^{\bar{\rho}_{k'(\lambda)}}_{\bar{\rho}_{k(\lambda)}} I^{\bar{\sigma}_{l'(1)}}_{\bar{\sigma}_{l(1)}} ...$$

$$= \sum_{k,l,..,k',l',..} \lambda_{k,l,..} \lambda_{k',l',..} (\nu(\rho))^{\mu_{(k' \cdot k^{-1})}} (\nu(\sigma))^{\mu_{(l' \cdot l^{-1})}} ...)$$

where $\mu_{(k' \cdot k^{-1})}$ is the number of cycles in the permutation k' followed by k^{-1}, etc. (The λ's are rational and therefore self-conjugate.) The number $\langle A \cdot A^* \rangle$ can thus be calculated and A vanishes if and only if this number is zero.

A 1.7d It will be seen from the argument given in A 1.7c above, that for any tensor system $\mathcal{S}$ over $\mathfrak{G}$, which is proper, the subsystem $\mathcal{R}$ is determined up to isomorphism by the (unordered) set of integers $\nu(\rho)$, $\nu(\sigma)$, ..., since these integers determine all the identities satisfied by the elements of $\mathcal{R}$. It will be convenient, in the case of such a proper system, to associate the symbol $\mathfrak{G}$ not only with semigroup of index types $\rho, \sigma, \ldots, \alpha, \ldots$, but also with the representation of this semigroup given by the positive integers $\nu(\rho), \nu(\sigma), \ldots, \nu(\alpha), \ldots$ With this interpretation of $\mathfrak{G}$, the proper system $\mathcal{R}$ over $\mathfrak{G}$ is essentially unique. This system will be denoted by $\mathcal{R}(\mathfrak{G})$.

A 1.8 Properties of $\mathcal{J}$. In A 1.5, the system $\mathcal{J}$ was introduced, this being a union of rings $\mathcal{J}_{(\lambda)}$ each of which is isomorphic with the group algebra of a symmetric group, over the field $\mathcal{F}_1$. Now $\mathcal{F}_1$ is without characteristic so that it is a splitting field for any symmetric group. Thus, each ring $\mathcal{J}_{(\lambda)}$ is a direct sum of rings $\mathcal{J}^{\{\lambda\}}$, each of which is isomorphic with a complete matrix algebra over $\mathcal{F}_1$. There is a ring corresponding to each partition λ of ν. (See van der Waerden (22) p.190.) In this subsection I shall employ a number of well-known results concerning these group-algebras. These may be found in Littlewood (11) and page references will be given where required.

A 1.8a The elements of the centre of the ring $\mathfrak{J}_{(\lambda)}$ are the elements of the form

$$\sum_{i=1}^{\lambda!} J_i \bullet A \bullet J_i^*$$

where $A \in \mathfrak{J}_{(\lambda)}$ and where the elements $J_1, \ldots, J_{\lambda!}$ are the permutations $J_{[k(1),\ldots,k(\lambda)]}$ (given in A 1.5h) in some order, with $J_1 = I = J_{[1,\ldots,\lambda]}$ and where, for each i, J_i^* is the permutation inverse to J_i. (This is in agreement with the notation of A 1.6p.) (See Littlewood (11) p.44)

A 1.8b The irreducible idempotents of the centre of $\mathfrak{J}_{(\lambda)}$ are the identity elements of the rings $\mathfrak{J}^{\{\lambda\}}$ for each partition λ of λ. The identity of $\mathfrak{J}^{\{\lambda\}}$ will be denoted by the symbol $T^{\{\lambda\}}$. The elements of the centre of $\mathfrak{J}^{\{\lambda\}}$ are then just the scalar multiples of $T^{\{\lambda\}}$.

We have the relations

$$\sum_{\lambda} T^{\{\lambda\}} = I^{(\lambda)} \qquad \text{A 1.8c}$$

$$T^{\{\lambda\}} \bullet T^{\{\mu\}} = \begin{cases} 0 \text{ if } \lambda \neq \mu \\ T^{\{\lambda\}} \text{ if } \lambda = \mu \end{cases} \qquad \text{A 1.8d}$$

$$T^{\{\lambda\}} \bullet A = A \bullet T^{\{\lambda\}} \qquad \text{A 1.8e}$$

for each $A \in \mathfrak{J}_{(\lambda)}$.

Denoting the group character corresponding to λ by the set of numbers $\chi_i^{(\lambda)}$, $(i = 1, \ldots, \lambda!)$, the elements $T^{\{\lambda\}}$ are given (See (11) p.46) by the formulae

$$\left.\begin{aligned} T^{\{\lambda\}} &= \frac{\chi_1^{(\lambda)}}{\lambda!} \sum_{i=1}^{\lambda!} \chi_i^{(\lambda)} J_i \\ &= \frac{\chi_1^{(\lambda)}}{\lambda!} \sum_{i=1}^{\lambda!} \chi_i^{(\lambda)} J_i^* \end{aligned}\right\} \quad \text{A 1.8f}$$

Each $\chi_i^{(\lambda)}$ is a rational number. $\chi_i^{(\lambda)}$ is the trace of the matrix representing J_i in the irreducible representation of $\mathfrak{J}_{(\lambda)}$ corresponding to the partition λ. Thus, $\chi_1^{(\lambda)}$ corresponds to the identity of $\mathfrak{J}_{(\lambda)}$.

A 1.8g Irreducible idempotents of the ring $\mathfrak{J}_{(\lambda)}$ may be obtained by the use of Young tableaux. Thus, if $\lambda = (\lambda_1, .., \lambda_p)$ is a partition of λ and $\mu = \tilde{\lambda} = (\mu_1, .., \mu_q)$ is the partition conjugate to λ, then a non-zero multiple of an irreducible idempotent of $\mathfrak{J}_{(\lambda)}$ may be obtained by choosing the permutation, $J^{\{\lambda\}}$ so that the expression

$$(N^{(\mu_1)} N^{(\mu_2)} \ldots N^{(\mu_q)}) \bullet (J^{\{\lambda\}} \bullet P^{(\lambda_1)} P^{(\lambda_2)} \ldots P^{(\lambda_p)} \bullet J^{\{\lambda\}*})$$

does not vanish. (The more usual form is to take the P's first rather than the N's.) Putting

$N = N^{(\mu_1)} \ldots N^{(\mu_q)}$, $P = P^{(\lambda_1)} \ldots P^{(\lambda_p)}$ and $J = J^{\{\lambda\}}$, we get

$$\kappa_{(\lambda)}(N \bullet J \bullet P \bullet J^*) \bullet (N \bullet J \bullet P \bullet J^*) = N \bullet J \bullet P \bullet J^*$$

for some scalar $\kappa_{(\lambda)}$ (see (11) p.72). Thus, $\kappa_{(\lambda)}\, N \bullet J \bullet P \bullet J^*$ is the required irreducible idempotent. In fact $\kappa_{(\lambda)}$ is rational and given by

$$\kappa_{(\lambda)} = \lambda_1! \ldots \lambda_p! \mu_1! \ldots \mu_q! \frac{\chi_1^{(\lambda)}}{\lambda!} .$$

Defining $K^{\{\lambda\}}$ and $K^{\{\lambda\}*}$ by

$$K^{\{\lambda\}} = N^{(\mu_1)} \dots N^{(\mu_q)} \cdot J^{\{\lambda\}} \cdot P^{(\lambda_1)} \dots P^{(\lambda_p)}$$

and A 1.8h

$$K^{\{\lambda\}*} = P^{(\lambda_1)} \dots P^{(\lambda_p)} \cdot J^{\{\lambda\}*} \cdot N^{(\mu_1)} \dots N^{(\mu_q)}$$

(see A 1.6q), we have

$$K^{\{\lambda\}} \cdot K^{\{\lambda\}*} \cdot K^{\{\lambda\}} \cdot J^* = (N \cdot J \cdot P) \cdot (P \cdot J^* \cdot N) \cdot (N \cdot J \cdot P) \cdot J^*$$

$$= N \cdot J \cdot P \cdot J^* \cdot N \cdot J \cdot P \cdot J^* = \frac{1}{\varkappa_{(\lambda)}} N \cdot J \cdot P \cdot J^*$$

by A 1.5r. Hence,

$$\varkappa_{(\lambda)} K^{\{\lambda\}} \cdot K^{\{\lambda\}*} \cdot K^{\{\lambda\}} = K^{\{\lambda\}} \qquad \text{A 1.8i}$$

Put

$$G^{\{\lambda\}} = \varkappa_{(\lambda)} K^{\{\lambda\}} \cdot K^{\{\lambda\}*}, \qquad \text{A 1.8j}$$

then $G^{\{\lambda\}} \cdot G^{\{\lambda\}} = G^{\{\lambda\}}$ and $G^{\{\lambda\}} \neq 0$ (since $G^{\{\lambda\}} \cdot K^{\{\lambda\}} = K^{\{\lambda\}}$), so that $G^{\{\lambda\}}$ is also an irreducible idempotent of $\underset{\sim}{\mathfrak{J}}_{(\lambda)}$. (It is irreducible since it contains an irreducible idempotent as a factor.)

We have

$$G^{\{\lambda\}} = P^{(\lambda)} \quad \text{and} \quad G^{\{1^\lambda\}} = N^{(\lambda)}$$

since $P^{(\lambda)}$ and $N^{(\lambda)}$ are, respectively, the unique irreducible idempotents of $\underset{\sim}{\mathfrak{J}}^{\{\lambda\}}$ and of $\underset{\sim}{\mathfrak{J}}^{\{1^\lambda\}}$.

A 1.8k Notation. If A is any element of $\underset{\sim}{\mathfrak{J}}_{(\lambda)}$, then the notation $\langle A \rangle_S$ will be used to denote the product of $\lambda!$ with the coefficient of the identity permutation in A. That is, if $A = \sum_{i=1}^{\lambda!} \lambda_i J_i$, then $\langle A \rangle_S = \lambda! \lambda_1$

This is the usual trace function for a group algebra (see (11) p.43).

For A, B $\in \mathfrak{J}_{(\lambda)}$, we have

A 1.8l $\langle A + B \rangle_S = \langle A \rangle_S + \langle B \rangle_S$,

A 1.8m $\langle A \cdot B \rangle_S = \sum \lambda_i \mu_i$, where $A = \sum \lambda_i J_i$, $B = \sum \mu_i J_i^*$,

A 1.8n $\langle A \cdot B \rangle_S = \langle B \cdot A \rangle_S$

A 1.8o $\langle T^{\{\lambda\}} \rangle_S = (\chi_1^{(\lambda)})^2$, by A 1.8f.

A 1.8p $\langle G^{\{\lambda\}} \rangle_S = \chi_1^{(\lambda)}$, since $G^{\{\lambda\}}$ is an irreducible idempotent (primitive characteristic unit) (See (11) p.56).

From A 1.8n and A 1.8p we have $\langle J_i \cdot G^{\{\lambda\}} \cdot J_i^* \rangle_S = \chi_1^{(\lambda)}$, for each permutation J_i . Also, since $G^{\{\lambda\}} \in \mathfrak{J}^{\{\lambda\}}$, it follows from A 1.8a and A 1.8b that $\sum J_i \cdot G^{\{\lambda\}} \cdot J_i^*$ is a scalar multiple of $T^{\{\lambda\}}$. Hence, by A 1.8o it follows that

$$T^{\{\lambda\}} = \frac{\chi_1^{(\lambda)}}{\lambda!} \sum_{i=1}^{\lambda!} J_i \cdot G^{\{\lambda\}} \cdot J_i^* . \qquad \text{A 1.8q}$$

A 1.8r Symmetric functions. Consider the set of symmetric functions of the n variables $\xi_1, .., \xi_n$. It is possible to construct a tensor system the scalars of which appear naturally as this set of symmetric functions. Let $\mathfrak{S}$ be generated by the single pair α , α^* with $\nu(\alpha) = n$. Let s be the diagonal matrix $s = \mathrm{diag}(\xi_1, .., \xi_n)$. Now any n × n matrix ay be considered as a tensor with one upper and one lower α -index. The operations of matrix addition and multiplication and the trace

operation for matrices can then be interpreted, respectively, as the tensor operations of addition, contracted product (A 1.3g) and trace (A 1.3m). Let $\mathcal{S}$ be the tensor system generated by the matrix $s = s^{\alpha}_{\alpha}$ and the elements of the system $\mathcal{R}(\mathfrak{S})$. Then the elements of $\mathcal{S}$ are of the form

$$\sum_{\substack{i,j \\ \rho_1+\ldots+\rho_t=r}} \lambda_{i,j,\rho_1,\ldots,\rho_t} J_i \bullet \{ (I_{(\alpha)})^{\rho_1} (s)^{\rho_2} (s \bullet s)^{\rho_3} \ldots (s \bullet s \bullet \ldots \bullet s)^{\rho_t} \} \bullet J_j$$

where the λ's are rational functions (with integer coefficients) of the scalars $\langle s \rangle$, $\langle s \bullet s \rangle$, ..., $\langle s \bullet s \bullet \ldots \bullet s \rangle$.

Now, the element $s \bullet s \bullet \ldots \bullet s$ (the contracted product of s with itself t times, say) is the matrix diag $(\xi_1^t , .. , \xi_n^t)$. Thus $\langle s \bullet s \bullet \ldots \bullet s \rangle = \xi_1^t + .. + \xi_n^t$. But any symmetric function of $\xi_1 , .. , \xi_n$ can be expressed in terms of such expressions. Hence, the scalars of $\mathcal{S}$ are just the (rational) symmetric functions of $\xi_1 , .. , \xi_n$.

A 1.8s Remark. If the matrix s of A 1.8r is replaced by any $n \times n$ matrix a^{α}_{α} whose eigenvalues are $\xi_1 , .. , \xi_n$ and which is diagonable by a similarity transformation, then the system $\mathcal{T}$ generated by a and the elements of $\mathcal{R}(\mathfrak{S})$ which is obtained is isomorphic with the system of A 1.8r and has the same set of scalars. However, if the matrix a is not diagonable, then $\mathcal{T}$ is not isomorphic with $\mathcal{S}$. In fact, it can be shown that a necessary and sufficient condition for the matrix a to be diagonable is that the system $\mathcal{T}$ be weakly regular (see A 1.6t).

A 1.8t Schur functions. Let s_α^α and $\mathcal{J}$ be as in A 1.8r. The Schur functions of $\xi_1, \ldots, \xi_n$ can then be written in a simple tensor form

Let $b_{\alpha_1 \ldots \alpha_r}$ be an ordinary tensor possessing r covariant indices. If a_α^α is a matrix of transformation, the transform b' of b is given by $b' = b \cdot (a)^r$. Now, b can be split into subtensors $b_{(i)}^{\{\lambda\}}$ each of which possesses the symmetry of an irreducible idempotent of $\underset{\sim}{\mathcal{J}}_{(r)}$. Thus, we may have

$$b_{(i)}^{\{\lambda\}} \cdot G^{\{\lambda\}} = b_{(i)}^{\{\lambda\}}$$

The transform of $b_{(i)}^{\{\lambda\}}$ given by the matrix a is

$$b_{(i)}^{\{\lambda\}} \cdot (a)^r = b_{(i)}^{\{\lambda\}} \cdot (G^{\{\lambda\}} \cdot (a)^r).$$

The expression $G^{\{\lambda\}} \cdot (a)^r$ can be regarded as the transforming matrix for $b_{(i)}^{\{\lambda\}}$. Hence, the Schur function of the eigenvalues of the matrix a, corresponding to the partition λ, is given by the trace $\langle G^{\{\lambda\}} \cdot (a)^r \rangle$ of this expression (see (11) p.188). Similarly, the Schur function of $\xi_1, \ldots, \xi_n$ corresponding to λ is $\langle G^{\{\lambda\}} \cdot (s)^r \rangle$.

A 1.8u Lemma. If $a_\alpha^\alpha \in \mathcal{J}_{\alpha\alpha^*}$ and $A \in \underset{\sim}{\mathcal{J}}_{(r)}$, then

$$A_{(\alpha)} \cdot (a)^r = (a)^r \cdot A_{(\alpha)}.$$

Proof. The result is true for permutations since $a_{\alpha_1}^{\alpha_{\kappa'(1)}} \ldots a_{\alpha_r}^{\alpha_{\kappa'(r)}} = a_{\alpha_{\kappa(1)}}^{\alpha_1} \ldots a_{\alpha_{\kappa(r)}}^{\alpha_r}$. Hence, by linearity, it is true generally.

A 1.8v Invariant matrices. Let a_α^α be an $n \times n$ matrix and let E be an idempotent element of $\underset{\sim}{\mathcal{J}}_{(r)}$; then the expression $E_{(\alpha)} \cdot (a)^r$ gives an

invariant matrix of a. For if $b = b^{\omega}_{\alpha}$, then

$$(E \bullet (a)^{\wedge}) \bullet (E \bullet (b)^{\wedge}) = E \bullet E \bullet (a)^{\wedge} \bullet (b)^{\wedge}$$
$$= E \bullet (a \bullet b)^{\wedge}$$

by A 1.8u and A 1.3k (see Littlewood (11) p.180).

If E is an irreducible idempotent, the matrix $E \bullet (a)^{\wedge}$ is essentially an irreducible invariant matrix. Thus, except for repeated or zero rows and columns, the $G^{\{\lambda\}} \bullet (a)^{\wedge}$ of A 1.8t is an irreducible invariant matrix. In particular, $N^{(\wedge)} \bullet (a)^{\wedge}$ and $P^{(\wedge)} \bullet (a)^{\wedge}$ are essentially the r^{th} compound matrix and the r^{th} induced matrix of a respectively (see (11) p.179).

A 1.8w Notation. Let A be any element of $\mathfrak{J}_{(\wedge)}$ and let s be as in A 1.8t ; then the expression $\langle A_{(\omega)} \bullet (s)^{\wedge} \rangle$ gives a symmetric function of $\xi_1, \ldots, \xi_n$. This symmetric function will be denoted by the symbol $\{A\}_S$ (or $\{A_{(\omega)}\}_S$).

By A 1.3n, A 1.3p and A 1.3k, A 1.3o and A 1.8u we have $\{A + B\}_S = \{A\}_S + \{B\}_S$, $\{AB\}_S = \{A\}_S \{B\}_S$ and $\{A \bullet B\}_S = \{B \bullet A\}_S$. The Schur function $\{\lambda\}$ is thus given by $\{\lambda\} = \{G^{\{\lambda\}}\}_S$.

The operation of plethysm (see (11) pp.206, 289) of Schur functions can also be written using this notation. Thus, if μ is a partition of t,

$$\{\lambda\} \otimes \{\mu\} = \{G^{\{\lambda\}}_{(\alpha)}\}_S \otimes \{G^{\{\mu\}}\}_S = \{(G^{\{\lambda\}}_{(\alpha)})^t \bullet G^{\{\mu\}}_{(\alpha^t)}\}_S$$
$$= \{G^{\{\mu\}}_{(\alpha^t)} \bullet (G^{\{\lambda\}}_{(\alpha)})^t\}_S$$

Since $\{J_i \cdot G^{\{\lambda\}} \cdot J_i^*\}_S = \{J_i^* \cdot J_i \cdot G^{\{\lambda\}}\}_S = \{G^{\{\lambda\}}\}_S$,

the following result follows from A 1.8q :

A 1.8x $\quad \{T^{\{\lambda\}}\}_S = \chi_1^{(\lambda)} \{G^{\{\lambda\}}\}_S$.

A 1.8y **Lemma.** If $\nu(\alpha) = n$ and λ is the partition $(\lambda_1, .., \lambda_p)$ of r, where $\lambda_1 \geqslant \lambda_2 \geqslant .. \geqslant \lambda_p$, then

$$\langle G^{\{\lambda\}}_{(\alpha)} \rangle = \frac{1}{\chi_1^{(\lambda)}} \langle T^{\{\lambda\}}_{(\alpha)} \rangle = \prod_{j=1}^{p} \prod_{i=1}^{\lambda_j} \frac{(n+i-j)}{(i+j-1)}$$

$$= \begin{vmatrix} \binom{n+\lambda_1-1}{\lambda_1} & \binom{n+\lambda_1}{\lambda_1+1} & \cdots & \binom{n+\lambda_1+p-2}{\lambda_1+p-1} \\ \binom{n+\lambda_2-2}{\lambda_2-1} & \binom{n+\lambda_2-1}{\lambda_2} & & \vdots \\ \vdots & & \ddots & \vdots \\ \binom{n+\lambda_p-p}{\lambda_p-p-1} & \cdots & \cdots & \binom{n+\lambda_p-1}{\lambda_p} \end{vmatrix}$$

Also, $\langle T^{\{\lambda\}}_{(\alpha)} \rangle = \frac{1}{n!} \prod_{j=1}^{p} \prod_{i=1}^{\lambda_j} (n+i-j)$.

Proof. In A 1.8r, if $\xi_1, .., \xi_n$ are all put equal to unity, the matrix s_α^α becomes the α-unit $I_{(\alpha)}$. Hence the expressions $G^{\{\lambda\}} \cdot (s)^\alpha$ and $T^{\{\lambda\}} \cdot (s)^\alpha$ become simply $G^{\{\lambda\}}$ and $T^{\{\lambda\}}$, respectively. The symmetric functions $\{G^{\{\lambda\}}\}_S$ and $\{T^{\{\lambda\}}\}_S$ thus specialize to $\langle G^{\{\lambda\}} \rangle$ and $\langle T^{\{\lambda\}} \rangle$. The determinantal formula for $\langle G^{\{\lambda\}}_{(\alpha)} \rangle$ is simply the Jacobi-Trudi formula $\{\lambda\} = |h_{\lambda_i - i + j}|$ for Schur functions (see (11) p.88) with $h_i = \binom{n+i-1}{i}$. The product formula is essentially that given by Littlewood for Schur functions for the series $\frac{1}{(1-x)^n}$. (See (11) pp.105 and 126. The formula on p.126 is, in fact,

misprinted but it is correctly given on p.105.) I have employed Littlewood's result $\chi_1^{(\lambda)} = \dfrac{n!}{\prod_{j=1}^{p}\prod_{i=1}^{\lambda_j}(i+j-1)}$.

The fact that $\langle G_{(\alpha)}^{\{\lambda\}}\rangle = \frac{1}{\chi_1^{(\alpha)}}\langle T_{(\alpha)}^{\{\lambda\}}\rangle$ follows from A 1.8x, or directly from A 1.8q.

A 1.9 Theorem. Let $\mathcal{J}$ be a proper tensor system over $\mathfrak{S}$ and let $\beta\in\mathfrak{S}$ with $\nu(\beta)=n$. Then any element A of $\mathcal{J}$ for which $A_{(\beta)}=0$ is of the form $A=\sum_{i,j}\zeta_{ij}\, J_i\bullet N^{(n+1)}I^{(i)}J_j$.

That is, the relation $N_{(\beta)}^{(n+1)}=0$ forms a basis for the relations satisfied by the elements of $\mathcal{J}^{(\beta)}$.

Proof. The elements of $\mathcal{J}$ which are mapped into the zero of $\mathcal{J}^{(\beta)}$ in the homomorphism $\Theta^{(\beta)}$, given by $B\to B_{(\beta)}$ (see A 1.5h), constitute a two-sided ideal in each algebra $\mathcal{J}_{(\lambda)}$. Now each $\mathcal{J}_{(\lambda)}$ is a direct sum of complete matrix algebras $\mathcal{J}^{\{\lambda\}}$ and the only two-sided ideals in a complete matrix algebra are the algebra itself and its zero sub-ideal. Thus, for each partition λ, either $\Theta^{(\beta)}$ maps the whole of $\mathcal{J}^{\{\lambda\}}$ into the zero of $\mathcal{J}^{(\beta)}$, or else it determines an isomorphism of $\mathcal{J}^{\{\lambda\}}$. Hence, if $0\neq A^{\{\lambda\}}\in\mathcal{J}^{\{\lambda\}}$ and $A_{(\beta)}^{\{\lambda\}}=0$, we must have $T_{(\beta)}^{\{\lambda\}}=0$ also. This implies $\langle T_{(\beta)}^{\{\lambda\}}\rangle=0$, so that, by A 1.8y, we have $\prod_{j=1}^{p}\prod_{i=1}^{\lambda_j}(n+i-j)=0$, where $\lambda=(\lambda_1,..,\lambda_p)$, $\lambda_1\geqslant..\geqslant\lambda_p$. Hence $p>n$. Now $A^{\{\lambda\}}=A^{\{\lambda\}}\bullet T^{\{\lambda\}}$ (see A 1.8b) and it follows from the formulae A 1.8q, A 1.8j and A 1.8h that $A^{\{\lambda\}}$ is of the form

CHAPTER 7

$$A^{\{\lambda\}} = \sum_{i,j} \zeta'_{ij}\ J_i \bullet N^{(\mu_1)}\ N^{(\mu_2)}\ \ldots\ N^{(\mu_q)} \bullet J_j\ ,$$

where $(\mu_1, \ldots, \mu_q)$ is the partition conjugate to λ. But $\mu_1 = p > n$, so that, by A 1.5r, $A^{\{\lambda\}}$ is of the form

$$A^{\{\lambda\}} = \sum_{i,j} \zeta''_{ij}\ J_i \bullet N^{(n+1)}\ I^{(\alpha)} \bullet J_j .$$

The result follows, since every element of $\mathfrak{J}$ is a sum of elements from the $\mathfrak{J}^{\{\lambda\}}$'s.

A 1.9a Remark. If the system $\mathcal{J}$ of theorem A 1.9 contains a β-unit but is not assumed to be proper, then other situations can arise. It can be shown that the only consistent possibilities are that either $\mathcal{J}^{(\beta)}$ is isomorphic with $\mathfrak{J}$, or else a base for relations satisfied by the elements of $\mathcal{J}^{(\beta)}$ is given by the vanishing of a Young tableau idempotent corresponding to a rectangular partition. If the basic identity is given by $K^{\{\lambda'\}}_{(\beta)} = 0$ (or, equivalently, $G^{\{\lambda'\}}_{(\beta)} = 0$), the index type β may be said to be of class $[s - 1, r - 1]$. Any index type α of a proper system is of class $[n, 0]$ where $n = \nu(\alpha)$. Only index types of class $[r, o]$ or $[o, s]$ can give rise to a normal system If β is of class $[r, s]$, then $\nu(\beta) = r - s$. Thus index types of class $[o, s]$ are of negative dimension ($s > o$). The consistency systems possessing index types of class $[o, s]$ will be indicated in B 3.11. An index type of class $[r, s]$ can be regarded as the direct sum of index types of classes $[r, o]$ and $[o, s]$ and hence the consistency of systems possessing index types of class $[r, s]$ may be

proved ($r \geqslant 0$, $s \geqslant 0$). The product of index types of respective classes $[r, s]$ and $[t, u]$ is one of class $[rt + su, st + ru]$.

A 1.9b **Definition.** The element $Q_m^{(\lambda)}$ of $\mathfrak{J}$ will be defined, for each $r > 0$, $n > 0$ by

$$Q_m^{(\lambda)} = \sum_{(\beta > m)} T^{\{\lambda\}} ,$$

the sum being taken over all partitions λ of r into more than n (non-zero) parts. If $n \leqslant r$, then $Q_n^{(\lambda)} = 0$.

Clearly, $Q_n^{(\lambda)}$ is an idempotent belonging to the centre of $\mathfrak{J}_{(\lambda)}$.

A 1.9c **Lemma.** If $\mathcal{S}$ is proper with $\nu(\beta) = n$ and $A \in \mathfrak{J}_{(r)}$, then $A_{(\beta)} = 0$ if and only if $A \cdot Q_m^{(\lambda)} = A$.

Proof. Since $Q_m^{(\lambda)}$ is a sum of $T^{\{\lambda\}}$'s each of which involves an $N^{(m+1)}$ (see proof of A 1.9), we have $Q_{m(\beta)}^{(\lambda)} = 0$ by A 1.7. Hence, if $A \cdot Q_m^{(\lambda)} = A$, then $A_{(\beta)} = A_{(\beta)} \cdot Q_{m(\beta)}^{(\lambda)} = 0$.

Conversely, if $A_{(\beta)} = 0$, then by the proof of A 1.9, A is a sum of elements $A^{\{\lambda\}}$ from the rings $\mathfrak{J}^{\{\lambda\}}$ for which the partition λ consists of more than n parts. The result follows since $A^{\{\lambda\}} \cdot T^{\{\lambda\}} = A^{\{\lambda\}}$ (and $A^{\{\lambda\}} \cdot T^{\{\mu\}} = 0$ if $\lambda \neq \mu$).

A 1.9d **Lemma.** Let $\mathcal{S}$ be proper with $\nu(\beta) = n$ and let $A \in \mathfrak{J}_{(\lambda)}$ where $A_{(\beta)} = 0$. Suppose A is of the form

$$A = B \cdot P^{(\lambda_1)} P^{(\lambda_2)} \ldots P^{(\lambda_\rho)}$$

CHAPTER 6

where $(\lambda_1, .., \lambda_p)$ is a partition of r with $p \leqslant n$; then $A = 0$.

Proof. By A 1.9c we have $A \bullet Q_m^{(n)} = A$. But $P^{(\lambda_1)} P^{(\lambda_2)} .. P^{(\lambda_p)} \bullet Q_m^{(n)} = 0$, since $Q_m^{(n)}$ involves $N^{(n+1)}$ with $n \geqslant p$ (see (10) p.74). The result follows.

A 2 Linear equations in a tensor system.

In this chapter the concept of a regular tensor system is introduced in analogy with that of a regular ring. Linear tensor equations are considered in such systems and in normal systems. For this purpose I shall introduce the idea of the generalized inverse for a tensor and of partial units in a tensor system. It is shown how a group algebra can be embedded in a tensor system so that the theory can be applied to linear equations in a group algebra.

A 2.1 Regular tensor systems

The tensor system $\mathcal{S}$ will be called regular if, for each element a_{β}^{α} of $\mathcal{S}$ there exists an element x_{α}^{β} of $\mathcal{S}$ satisfying

$$a \bullet x \bullet a = a.$$

The use of the word regular, in this sense, is in line with that of von Neumann (15) for the same condition in a ring. It will be seen that any regular tensor system is also weakly regular (see A 1.6t), since

if $\alpha = 1$ in the above, we have that for any $a_\beta \in \mathcal{S}, a \neq 0$, there exists an $x^\beta \in \mathcal{S}$ such that $a_\beta x^\beta = 1 \neq 0$.

Conversely, there is the following result :

A 2.1a **Lemma.** Let $\mathcal{S}$ be a weakly regular tensor system over $\mathfrak{S}$. Suppose that the index type $\alpha \in \mathfrak{S}$ has the property for some $r > 0$, that for every $b^{\gamma}_{\alpha_1 \cdots \alpha_\lambda} \in \mathcal{S}_{\alpha \wedge \gamma^*}$ there is a linear relation connecting the transposes of b obtained by permuting its indices $\alpha_1, \ldots, \alpha_\lambda$. Then, for each $a^{\beta}_{\alpha} \in \mathcal{S}_{\alpha\beta^*}$, there is an $x^{\alpha}_{\beta} \in \mathcal{S}_{\alpha^*\beta}$ satisfying $a \bullet x \bullet a = a$.

Proof. The element a satisfies a relation of the form

$$\sum_k \mu_k \, a^{\beta_1}_{\alpha_{k(1)}} \cdots a^{\beta_\lambda}_{\alpha_{k(\lambda)}} = 0,$$

summed over all permutations k of $1, \ldots, r$. We may suppose that r is as small as possible, given a, and also that the coefficient μ_1 corresponding to the identity permutation is non-zero. (Otherwise a transpose of this expression could have been taken instead.) This may be written

$$w^{\beta_1 \cdots \beta_{\lambda-1}}_{\alpha_1 \cdots \alpha_{\lambda-1}} \, a^{\beta_\lambda}_{\alpha_\lambda} = \sum \mu_k \, a^{\beta_1}_{\alpha_{k(1)}} \cdots a^{\beta_\lambda}_{\alpha_{k(\lambda)}}$$

where the sum is taken over all permutations k for which $k(r) \neq r$. Now w is non-zero, since $\mu_1 \neq 0$ and r was chosen to be as small as possible. Hence there exists $y^{\alpha_{\lambda-1}}_{\beta_{\lambda-1}} \in \mathcal{S}$ such that $\langle w \bullet y \rangle = 1$, $\mathcal{S}$ being weakly regular. Thus,

$$a = \langle w \bullet y \rangle \, a = \sum_{k(\lambda) \neq \lambda} a^{\bar\beta_1}_{\bar\alpha_{k(1)}} \cdots a^{\bar\beta_{k^{-1}(\lambda)}}_{\alpha} \cdots a^{\bar\beta_{\lambda-1}}_{\bar\alpha_{k(\lambda-1)}} \, a^{\beta}_{\bar\alpha_{k(\lambda)}} \, y^{\bar\alpha_1 \cdots \bar\alpha_{\lambda-1}}_{\bar\beta_1 \cdots \bar\beta_{\lambda-1}},$$

whence $a = a_{\alpha}^{\bar{\beta}}\, a_{\bar{\gamma}}^{\beta}\, x_{\bar{\beta}}^{\bar{\alpha}} = a \bullet x \bullet a$ for suitable x_{β}^{α} in $\mathfrak{S}$.

A 2.1b Lemma. Any proper tensor system which is weakly regular is also regular.

Proof. This follows from A 1.7 and A 2.1a (see A 1.5j) with $r = \nu(\alpha) + 1$, for each α.

A 2.1c Henceforth, all tensor systems considered will be supposed to be proper tensor systems unless it is explicitly stated to the contrary. Thus, by A 2.1b, all weakly regular systems and, in particular, all normal systems will be regular.

A 2.2 Theorem. Let $\mathfrak{S}$ be a normal tensor system ; then if $a_{\beta}^{\alpha} \in \mathfrak{S}$, the four equations

$$a \bullet x \bullet a = a \qquad \text{A 2.2a}$$

$$x \bullet a \bullet x = x \qquad \text{A 2.2b}$$

$$(a \bullet x)^{*} = a \bullet x \qquad \text{A 2.2c}$$

$$(x \bullet a)^{*} = x \bullet a \qquad \text{A 2.2d}$$

have a unique solution in $\mathfrak{S}$.

Proof. It will first be shown that equations A 2.2b and A 2.2c are equivalent to

$$x \bullet x^{*} \bullet a^{*} = x \,. \qquad \text{A 2.2e}$$

Now, equation A 2.2e follows from A 2.2b and A 2.2c, since it is merely A 2.2c substituted in A 2.2b (by A 1.6b). Conversely, A 2.2e implies $a \bullet x \bullet x^{*} \bullet a^{*} = a \bullet x$, the left-hand side of which is self-conjugate. Thus

A 2.2c follows: and, substituting A 2.2c in A 2.2e, we get A 2.2b.

Similarly, A 2.2a and A2.2d can be replaced by the equation

$$x \cdot a \cdot a^* = a^* \qquad \text{A 2.2f}$$

Thus it is sufficient to find an x satisfying A 2.2e and A 2.2f. Such an x will exist if an element u of $\mathcal{S}$ can be found satisfying

$$u \cdot a^* \cdot a \cdot a^* = a^* . \qquad \text{A 2.2g}$$

For then $x = u \cdot a^*$ satisfies A 2.2f. Also, we have seen that A 2.2f implies $a^* \cdot x^* \cdot a^* = a^*$ and therefore $u \cdot a^* \cdot x^* \cdot a^* = u \cdot a^*$.

Thus x also satisfies A 2.2e.

Now $\mathcal{S}$ is regular (see A 2.1c) so there exists an element v of satisfying

$$a^* \cdot a \cdot a^* \cdot v \cdot a^* \cdot a \cdot a^* = a^* \cdot a \cdot a^* .$$

Applying A 1.6m once, and then again with a^* replacing a, we get

$$a^* \cdot v \cdot a^* \cdot a \cdot a^* = a^* .$$

Thus, if $u = a^* \cdot v$, A 2.2g is satisfied.

To show that x is unique, we suppose that x satisfies A 2.2e and A 2.2f and that y satisfies $y = a^* \cdot y^* \cdot y$ and $a^* = a^* \cdot a \cdot y$. These last relations are obtained by respectively substituting A 2.2d in A 2.2b and A 2.2c in A 2.2a. (They are A 2.2e and A 2.2f with y in place of x and the reverse order of multiplication and must, by symmetry, also be equivalent to A 2.2a, ... A 2.2d.) Now

$$x = x \cdot x^* \cdot a^* = x \cdot x^* \cdot a^* \cdot a \cdot y = x \cdot a \cdot y = x \cdot a \cdot a^* \cdot y^* \cdot y = a^* \cdot y^* \cdot y = y.$$

A 2.2h The unique solution of A 2.2a, .., A 2.2d will be called the generalized inverse of a and will be written $a^\dagger$. The definition of $a^\dagger$ depends (as does the notation a•b, etc.) on the particular division of the indices of a into upper and lower indices. Thus any element of $\mathcal{S}$ has a number of generalized inverses, one for each of the different ways of dividing the indices into two groups.

The existence of such a generalized inverse for ordinary matrices is discussed in Penrose (17). The same operation was also described by Moore (13), (14) but defined in a somewhat different way (see Rado (19)).

The following formulae are easily deduced from the definition of the generalized inverse :

A 2.2i $\quad (a^\dagger)^* = (a^*)^\dagger,$

A 2.2j $\quad (a^\dagger)^\dagger = a,$

A 2.2k if a^{α}_{β} is such that there exists b^{β}_{α} satisfying $a \bullet b = I_{(\alpha)}$ and $b \bullet a = I_{(\beta)}$, then $b = a^\dagger$,

A 2.2l $\quad a^\dagger = (a^* \bullet a)^\dagger \bullet a^*,$

A 2.2m $\quad (ab)^\dagger = a^\dagger b^\dagger,$

(but note that $(a \bullet b)^\dagger \neq b^\dagger \bullet a^\dagger$, in general)

A 2.2n $\quad a^* = a$ implies $a \bullet a^\dagger = a^\dagger \bullet a,$

A 2.2o if λ is a scalar with $\lambda^* = \lambda$ and $\lambda\, a \bullet a^* \bullet a = a$, then $a^\dagger = \lambda\, a^*$.

A 2.3 Rank of a tensor.

I shall use the term rank in its matrix sense and not in the sense that it is commonly applied to tensors (i.e. not in the sense that any tensor of type $\rho^{\sim}$, say, might be called a tensor of rank r).

A 2.3a Definition. Let $\mathcal{S}$ be a tensor system and let $a_{\alpha}^{\beta} \in \mathcal{S}$; then a will be said to be of rank r if

$$(a)^{\lambda+1} \cdot N^{(\lambda+1)} = 0 \text{ and } (a)^{\lambda} \cdot N^{(\lambda)} \neq 0.$$

This definition,also,depends on the division of the indices into two groups.

Since $N_{(\alpha)}^{(\nu(\alpha)+1)} = 0$ ($\mathcal{S}$ being proper, see A 2.1c) it follows that every element of $\mathcal{S}$ has a rank. The zeros of $\mathcal{S}$ have rank zero because $(0)^{0}$ is defined as unity (see A 1.3s).

A 2.3b If a_{α}^{β} is an ordinary matrix, then $(a)^{\lambda} \cdot N^{(\lambda)}$ $(= N_{(\beta)}^{(\alpha)} \cdot (a)^{\lambda})$ is essentially the s^{th} compound matrix of a (see A 1.8v). Now, if a matrix has rank r, its $(r+1)^{th}$ compound vanishes whereas its r^{th} compound does not. Hence, the definition A 2.3a agrees with the definition of rank for matrices.

A 2.3c Lemma. If $a_{\alpha}^{\alpha} \in \mathcal{S}$ has rank n and $a \cdot a = a$, then $\langle a \rangle = n$.

Proof. This follows by an argument exactly similar to that of A 1.7 but with a replacing the α-unit $I_{(\alpha)}$ and $\langle a \rangle$ replacing $\nu(\alpha)$.

It is a well known property for an ordinary matrix that, if it is idempotent, then its rank is equal to its trace.

A 2.3d **Lemma.** Let $\mathcal{S}$ be normal and let $a_\alpha^\beta \in \mathcal{S}$ have rank r. Then

$$a^\dagger = \frac{r}{\langle N^{(r)} \bullet (a^* \bullet a)^r \rangle} \left(N^{(r)} \bullet a^* (a^* \bullet a)^{r-1} \right)_{[\alpha^{r-1}]} .$$

Proof. We have $N^{(r+1)} \bullet (a^*)^{r+1} = 0$ by 1.61, etc. Thus, by A 1.5t (with indices in the reverse order), we get

$$\frac{1}{r+1} a^* \left(N^{(r)} \bullet (a^*)^r \right) = \frac{r}{r+1} N^{\alpha_2 \cdots \alpha_{r+1}}_{\bar\alpha_1 \cdots \bar\alpha_r} N^{\alpha_1 \bar\alpha_2 \cdots \bar\alpha_r}_{\bar\alpha_{r+1} \cdots \bar\alpha_{2r}} a^{*\,\bar\alpha_1}_{\beta_1} a^{*\,\bar\alpha_{r+1}}_{\beta_2} \cdots a^{*\,\bar\alpha_{2r}}_{\beta_{r+1}} .$$

Hence

$$\begin{aligned} a^* \langle N^{(r)} \bullet (a^*)^r \bullet (a)^r \rangle &= r\, N^{\bar\alpha_{2r+1} \cdots \bar\alpha_{3r+1}}_{\bar\alpha_1 \cdots \bar\alpha_r} N^{\alpha \bar\alpha_2 \cdots \bar\alpha_r}_{\bar\alpha_{r+1} \cdots \bar\alpha_{2r}} a^{*\,\bar\alpha_1}_{\beta} a^{*\,\bar\alpha_{r+1}}_{\bar\beta_1} \cdots a^{*\,\bar\alpha_{2r}}_{\bar\beta_r} a^{\bar\beta_1}_{\bar\alpha_{2r+1}} \cdots a^{\bar\beta_r}_{\bar\alpha_{3r+1}} \\ &= r \left\{ \left(N^{(r)} \bullet (a^* \bullet a)^r \bullet N^{(r)} \right)_{[\alpha^{r-1}]} \right\} \bullet a^* \\ &= r \left\{ \left(N^{(r)} \bullet (a^* \bullet a)^r \right)_{[\alpha^{r-1}]} \right\} \bullet a^* \\ &= r \left\{ \left(N^{(r)} \bullet I(a^* \bullet a)^{r-1} \right)_{[\alpha^{r-1}]} \right\} \bullet a^* \bullet a \bullet a^* \end{aligned}$$

by A 1.8u and A 1.8r. Now $(a)^r \bullet N^{(r)} \neq 0$ so, by A 1.6j, etc.,

$$0 \neq \langle ((a)^r \bullet N^{(r)})^* \bullet ((a)^r \bullet N^{(r)}) \rangle = \langle N^{(r)} \bullet (a^* \bullet a)^r \rangle = \lambda$$

say. Thus, putting $u = \frac{r}{\lambda} \left(N^{(r)} \bullet I(a^* \bullet a)^{r-1} \right)_{[\alpha^{r-1}]}$, we get $a^* = u \bullet a^* \bullet a \bullet a^*$. This is relation A 2.2g, so that $a^\dagger = u \bullet a^*$ as required.

A 2.3e **Lemma.** If $\mathcal{S}$ is regular and $a_\alpha^\beta \in \mathcal{S}$ has rank r, then a solution of $a \bullet x \bullet a = a$ is given by $x = \frac{r}{\mu} \left(N^{(r)} \bullet b \bullet N^{(r)} \bullet I(a)^{r-1} \right)_{[\alpha^{r-1}]}$ where b is chosen so that $0 \neq \mu = \langle N^{(r)} \bullet b \bullet N^{(r)} \bullet (a)^r \rangle$.

Proof. The proof is similar to that of A 2.3d. This is the result which would have been obtained if the proof of A 2.1a had been carried out explicitly for an element a of rank r.

A 2.4 Theorem. Let $\mathcal{S}$ be normal and let $a_\alpha^\beta \in \mathcal{S}$ have rank r. Let $c_{(1)\alpha}^{\ \ \alpha}$, $c_{(2)\alpha}^{\ \ \alpha}$, .. be defined inductively by

$$c_{(1)} = I_{(\alpha)} \quad \text{and} \quad c_{(i+1)} = \frac{1}{i}\left\langle c_{(i)} \cdot a^* \cdot a \right\rangle I - c_{(i)} \cdot a^* \cdot a$$

$(i = 1, 2, ..)$, then $c_{(r+1)} \cdot a^* \cdot a = 0$ and

$$a^\dagger = \frac{r}{\left\langle c_{(r)} \cdot a^* \cdot a \right\rangle} c_{(r)} \cdot a \ .$$

Proof. Put $d_{(i)} = i\left(N^{(i)} \cdot I(a^* \cdot a)^{i-1}\right)_{[\alpha^{i-1}]}$, $(i = 1, 2, ..)$.

Then, applying A 1.5t (with indices reversed) to $d_{(i+1)}$, we obtain

$$d_{(i+1)} = I\left\langle N^{(i)} \cdot (a^* \cdot a)^i \right\rangle - i N^{\alpha \bar\alpha_1 \cdots \bar\alpha_{2i-1}}_{\alpha \bar\alpha_1 \cdots \bar\alpha_{i-1}} \ N^{\bar\alpha_1 \bar\alpha_2 \cdots \bar\alpha_{2i-1}}_{\bar\alpha_2 \bar\alpha_3 \cdots \bar\alpha_{3i-1}} a^{*\bar\alpha_{2i}}_{\beta_1} \cdots a^{*\bar\alpha_{3i-1}}_{\beta_{i-1}} a^{\beta_i}_{\bar\alpha_i} \cdots a^{\beta_{i-1}}_{\bar\alpha_{2i-1}} .$$

Thus, similarly to the proof of A 2.3d, we get $d_{(i+1)}$

$= \frac{1}{i}\left\langle d_{(i)} \cdot a^* \cdot a \right\rangle I - d_{(i)} \cdot a^* \cdot a$, whence $d_{(i)} = c_{(i)}$ $(i = 1, 2, ..)$.

Hence, by A 2.3d, $a^\dagger = \frac{r}{\left\langle c_{(r)} \cdot a^* \cdot a \right\rangle} c_{(r)} \cdot a^*$ so that $c_{(r+1)} \cdot a^* \cdot a = 0$.

A 2.4a The result A 2.4 gives an algorithm for computing the generalized inverse of any tensor belonging to a normal system. The corresponding result for ordinary matrices is given in Penrose (19). If $a^* \cdot a$ is replaced by any non-singular matrix b, we get $b^{-1} = \frac{r}{\left\langle c_{(r)} \cdot b \right\rangle} c_{(r)}$. This is Frame's method for computing the inverse of a matrix (see Dwyer (4)).

A 2.4b With the notation of A 2.4, we have $I - a^\dagger \cdot a = \frac{r}{\left\langle c_{(r)} \cdot a^* \cdot a \right\rangle} c_{(r+1)}$, as follows easily from A 2.4. Hence,

$$I - a^\dagger \cdot a = \frac{r}{\left\langle N^{(r)} \cdot (a^* \cdot a)^r \right\rangle} \left(N^{(r+1)} \cdot I(a^* \cdot a)^r\right)_{[\alpha^r]} .$$

A 2.5 Theorem. Let a^{α}_{β}, c^{α} be elements of a normal system $\mathcal{S}$. Then a necessary and sufficient condition for the equation $a \bullet x = c$ to have a solution in $\mathcal{S}$ is $a \bullet a^{\dagger} \bullet c = c$, in which case the general solution is $x = a^{\dagger} \bullet c + y - a^{\dagger} \bullet a \bullet y$, where y^{β} is an arbitrary element of $\mathcal{S}_{\beta *}$.

Proof. Suppose $x^{\beta} \in \mathcal{S}$ satisfies $a \bullet x = c$. Then $a \bullet a^{\dagger} \bullet c = a \bullet a^{\dagger} \bullet a \bullet x = a \bullet x = c$. Conversely, suppose $a \bullet a^{\dagger} \bullet c = c$. Then $x = a^{\dagger} \bullet c + y - a^{\dagger} \bullet a \bullet y$ is a solution of $a \bullet x = c$, for any element y^{β} of $\mathcal{S}_{\beta *}$. Also, if x satisfies $a \bullet x = c$, then $x = a^{\dagger} \bullet c + x - a^{\dagger} \bullet a \bullet x$, proving the result.

For the case of ordinary matrices, this result is given in Moore (13) p.209 and Penrose (17) p.409.

A 2.5a If in A 2.5 the system $\mathcal{S}$ is supposed only to be regular, the element $a^{\dagger}$ may be replaced by any element of $\mathcal{S}$ satisfying $a \bullet x \bullet a = a$ (e.g. sec A 2.3e). The proof is exactly similar.

A 2.5b The equation $a \bullet x = c$ considered in A 2.5 is the most general type of linear equation in a tensor system. In a non-commutive ring, the most general type of linear equation is of the form

$$a_{(1)} \bullet x \bullet b_{(1)} + \ldots + a_{(\lambda)} \bullet x \bullet b_{(\lambda)} = c.$$

However, if $a_{(1)\beta}^{\alpha}$, .., $a_{(\lambda)\beta}^{\alpha}$, $b_{(1)\delta}^{\gamma}$, .., $b_{(\lambda)\delta}^{\gamma}$, c_{δ}^{α} are elements of a tensor system $\mathcal{S}$, this equation can be written in the form $d \bullet x' = c'$, where $d^{\alpha\delta *}_{\beta\gamma *} = a_{(1)\beta}^{\alpha} b_{(1)\delta}^{\gamma} + .. + a_{(\lambda)\beta}^{\alpha} b_{(\lambda)\delta}^{\gamma}$, $x'^{\beta\gamma *} = x^{\beta}_{\gamma}$ and $c'^{\alpha\delta *} = c^{\alpha}_{\delta}$.

In particular, when r = 1 and $\mathcal{J}$ is normal we have, by A 2.2m,

A 2.5c <u>Corollary</u>. If $\mathcal{J}$ is normal and a^{α}_{β}, b^{γ}_{δ}, $c^{\alpha}_{\delta} \in \mathcal{J}$, the general solution of the equation $a \bullet x \bullet b = c$ is $x = a^{\dagger} \bullet c \bullet b^{\dagger} + y - a^{\dagger} \bullet a \bullet y \bullet b \bullet b^{\dagger}$, provided that the equation has a solution, where y^{γ}_{β} is an arbitrary element of $\mathcal{J}_{\gamma * \beta}$.

Simultaneous linear equations will be considered in A 7.5.

A 2.6 <u>Theorem</u>. Let a^{α}_{β}, $c^{\alpha}_{\gamma} \in \mathcal{J}$. Then a necessary condition for the equation $a \bullet x = c$ to have a solution in $\mathcal{J}$ is the existence of an integer r such that $N^{(r)} \bullet (a)^{r} \neq 0$ and $N^{(r+1)} \bullet c(a)^{r} = 0$.

Conversely, if $\mathcal{J}$ is regular, the condition is also sufficient.

<u>Proof</u>. First, suppose x is an element of $\mathcal{J}$ satisfying $a \bullet x = c$. Then if r is the rank of a, we have $N^{(r)} \bullet (a)^{r} \neq 0$ and $N^{(r+1)} \bullet (a)^{r+1} = 0$. Hence, $N^{(r+1)} \bullet (a \bullet x)(a)^{r} = 0$; that is, $N^{(r+1)} \bullet c(a)^{r} = 0$. For the converse, suppose that there is an integer r such that $N^{(r)} \bullet (a)^{r} \neq 0$ and $N^{(r+1)} \bullet c(a)^{r} = 0$. If $\mathcal{J}$ is regular, it contains an element $b^{\beta^{r}}_{\alpha^{r}}$ such that $\langle b \bullet N^{(r)} \bullet (a)^{r} \rangle = 1$. Also, applying A 1.8t to $N^{(r+1)} \bullet c(a)^{r} = 0$, we get

$$c\{N^{(r)} \bullet (a)^{r}\} = N^{\alpha_2 \cdots \alpha_{r+1}}_{\bar{\alpha}_1 \cdots \bar{\alpha}_r} N^{\alpha_1 \bar{\alpha}_2 \cdots \bar{\alpha}_r}_{\bar{\alpha}_{r+1} \cdots \bar{\alpha}_{2r}} c^{\bar{\alpha}_1}_{\gamma} a^{\bar{\alpha}_{r+1}}_{\beta_1} \cdots a^{\bar{\alpha}_{2r}}_{\beta_r}$$

$$= a^{\alpha_1}_{\bar{\beta}_1} \left\{ N^{\alpha_2 \cdots \alpha_{r+1}}_{\bar{\alpha}_1 \cdots \bar{\alpha}_r} c^{\bar{\alpha}_1}_{\gamma} a^{\bar{\alpha}_2}_{\bar{\beta}_2} \cdots a^{\bar{\alpha}_r}_{\bar{\beta}_r} N^{\bar{\beta}_1 \cdots \bar{\beta}_r}_{\beta_1 \cdots \beta_r} \right\}$$

by A 1.8u. Hence contracting with $b^{\beta^{r}}_{\alpha^{r}}$ we obtain a relation of the form $c = a \bullet x$ with $x \in \mathcal{J}$.

A 2.6a If a and c are ordinary matrices, the classical condition for the existence of a solution of the equation $a \bullet x = c$ is that the rank of the matrix a be equal to the rank of the matrix a bordered by c. Theorem A 2.6 gives a tensor equivalent of this.

A 2.6b It is also possible to relate A 2.6 to the condition given in A 2.5 for $a \bullet x = c$ to have a solution. For let $\mathcal{S}$ be normal. The condition for solution according to A 2.5 is $(I - a \bullet a^{\dagger}) \bullet c = 0$. If r is the rank of a, this is equivalent, by A 2.4b, to $(N^{(r+1)} \bullet I(a \bullet a^*)^r)_{[\alpha^r]} \bullet c = 0$, i.e. $(I(a^*)^r \bullet N^{(r+1)*} \bullet N^{(r+1)} \bullet I(a)^r)_{[\alpha^r]} \bullet c = 0$, which is equivalent, by A 1.6m, to $N^{(r+1)} \bullet c(a)^r = 0$.

The solution to $a \bullet x = c$, given in the proof of A 2.6, is that which would be obtained by applying A 2.3e to A 2.5a.

A 2.6c Theorem A 2.6 shows that if $\mathcal{S}$ is regular, the equation $a \bullet x = c$ has a solution in $\mathcal{S}$ if an only if it has a solution in some extension of $\mathcal{S}$. Conversely, if $\mathcal{S}$ is a subsystem of some regular system $\mathcal{T}$, this property holds only if $\mathcal{S}$ is regular, since c can be taken as the unit of $\mathcal{S}_1$.

A 2.7 Theorem. Let $a_{(1)\alpha}, \ldots, a_{(r)\alpha}$ be r elements of $\mathcal{S}_\alpha$. Then a necessary condition for there to exist a relation

$$\lambda_1 a_{(1)} + \ldots + \lambda_r a_{(r)} = 0$$

for some $\lambda_1, \ldots, \lambda_r \in \mathcal{S}_1$ not all zero, is that

$$a_{(1)} \ldots a_{(r)} \bullet N^{(r)} = 0 .$$

Conversely, if $\mathcal{S}$ is regular, the condition is also sufficient.

Proof. Suppose $\lambda_1 a_{(1)} + .. + \lambda_r a_{(r)} = 0$ where $\lambda_1 \neq 0$, say. Then

$$a_{(1)} .. a_{(r)} \bullet N^{(r)} = -\frac{1}{\lambda_1}\{\lambda_2 a_{(2)} + .. + \lambda_r a_{(r)}\} a_{(2)} .. a_{(r)} \bullet N^{(r)}$$

$$= -\frac{\lambda_2}{\lambda_1} a_{(2)} a_{(2)} .. a_{(r)} \bullet N^{(r)} - .. - \frac{\lambda_r}{\lambda_1} a_{(r)} a_{(2)} .. a_{(r)} \bullet N^{(r)} = 0.$$

(All terms vanish by the skew-symmetry of $N^{(r)}$.) For the converse, suppose, as an inductive hypothesis, that $a_{(1)} .. a_{(r-1)} \bullet N^{(r-1)} = 0$ implies the existence of scalars $\mu_1, .., \mu_{r-1}$, not all zero, for which $\mu_1 a_{(1)} + .. + \mu_{r-1} a_{(r-1)} = 0$. (This is trivial for r = 2.) Now suppose $a_{(1)} .. a_{(r)} \bullet N^{(r)} = 0$. We have

$$0 = a_{(1)} .. a_{(r)} \bullet N^{(r)} = \frac{1}{r}\sum_{j=1}^{\hat{r}} (-1)^{r-j} (a_{(1)} .. a_{(j-1)} a_{(j+1)} .. a_{(r)} \bullet N^{(r-1)}) a_{(j)}$$

by A 1.5s. Either $a_{(1)} .. a_{(r-1)} \bullet N^{(r-1)} = 0$, in which case the required linear relation is given by $\mu_1 a_{(1)} + .. + \mu_{r-1} a_{(r-1)} + 0\, a_{(r)} = 0$, or else there exists $b^{\alpha^{r-1}} \in \mathcal{S}$ such that $a_{(1)} .. a_{(r-1)} \bullet N^{(r-1)} \bullet b \neq 0$, provided that $\mathcal{S}$ is regular. In this case, if

$$\lambda_{(j)} = \frac{(-1)^{r-j}}{r} a_{(1)} .. a_{(j-1)} a_{(j+1)} .. a_{(r)} \bullet N^{(r-1)} \bullet b$$

we have $\lambda_r \neq 0$ and $\lambda_1 a_{(1)} + .. + \lambda_r a_{(r)} = 0$, as required.

2.7a If $\mathcal{S}$ is a system of ordinary tensors over a field $\mathcal{F}_1$, then, by A 1.5m,

$$a_{(1)} .. a_{(r)} \bullet N^{(r)} = \begin{vmatrix} a_{(1)\alpha_1} & - - & a_{(r)\alpha_1} \\ \vdots & & \vdots \\ a_{(1)\alpha_r} & - - & a_{(r)\alpha_r} \end{vmatrix} .$$

CHAPTER 6

Thus A 2.7 gives the usual condition for linear dependence of vectors.

A 2.7b Lemma. If $\mathfrak{F}$ is regular, each vector space $\mathfrak{F}_\alpha$ is of finite dimension.

Proof. By A 1.7, $N_{(\alpha)}^{(\nu(\alpha)+1)} = 0$. Hence, by A 2.7, any $\nu(\alpha)+1$ elements of $\mathfrak{F}_\alpha$ must be linearly dependent over $\mathfrak{F}_1$.

A 2.7c Lemma. Let $\mathfrak{F}$ be regular and let $x_{(1)\alpha}, \ldots, x_{(\lambda)\alpha}$ be r linearly independent elements of $\mathfrak{F}$ which span $\mathfrak{F}_\alpha$. Then there exists a unique set, $y^\alpha_{(1)}, \ldots, y^\alpha_{(\lambda)}$, of r linearly independent elements of $\mathfrak{F}$ which span $\mathfrak{F}_{\alpha^*}$ for which $x_{(i)} \bullet y_{(j)} = \delta_{ij}$ $(i, j = 1, \ldots, r)$.

Proof. By A 2.7, $x_{(1)} \ldots x_{(\lambda)} \bullet N^{(\lambda)} \neq 0$. Thus, since $\mathfrak{F}$ is regular, there exists $a^{\alpha\wedge} \in \mathfrak{F}$ such that $x_{(1)} \ldots x_{(\lambda)} \bullet N^{(\lambda)} \bullet a = 1$. Put

$$y_{(j)} = x_{(1)} \ldots x_{(j-1)}\, I\, x_{(j+1)} \ldots x_{(\lambda)} \bullet N^{(\lambda)} \bullet a .$$

Then $x_{(i)} \bullet y_{(j)} = x_{(1)} \ldots x_{(j-1)}\, x_{(i)}\, x_{(j+1)} \ldots x_{(\lambda)} \bullet N^{(\lambda)} \bullet a = \delta_{ij}$, by the skew-symmetry of $N^{(\lambda)}$. The $y_{(j)}$ are linearly independent since $\sum_j \lambda_j y_{(j)} = 0$ implies $0 = \sum_j \lambda_j x_{(i)} \bullet y_{(j)} = \lambda_i$ $(i = 1, \ldots, r)$. Hence, the dimension of $\mathfrak{F}_{\alpha^*}$ is greater than or equal to that of $\mathfrak{F}_\alpha$. Since $(\alpha^*)^* = \alpha$ it follows that the dimension of $\mathfrak{F}_\alpha$ is greater than or equal to that of $\mathfrak{F}_{\alpha^*}$. Thus the dimensions are equal and $y_{(1)}, \ldots, y_{(\lambda)}$ span $\mathfrak{F}_{\alpha^*}$. The uniqueness of $y_{(1)}, \ldots, y_{(\lambda)}$ follows, since if $x_{(i)} \bullet y'_{(j)} = \delta_{ij}$ with $y'_{(j)} \in \mathfrak{F}_{\alpha^*}$ $(i, j = 1, \ldots, r)$, we have $y'_{(j)} = \sum_k \mu_{jk} y_{(k)}$, whence $\delta_{ij} = \sum_k \mu_{jk} \delta_{ik}$ so that $\mu_{ji} = \delta_{ij}$ and $y'_{(j)} = y_{(j)}$.

A 2.7d Remark. If $\mathcal{S}$ is normal it is possible, for each α, to choose the $x_{(i)\alpha}$ and $y^{\alpha}_{(i)}$ of A 2.7c so that $y_{(i)} = \lambda_i \, x^{*}_{(i)}$ for some $\lambda_i \in \mathcal{S}_1$ $(i = 1, .., r)$. If $z_{(1)\alpha}$, .., $z_{(\lambda)\alpha}$ span $\mathcal{S}_\alpha$, this may be done by putting $x_{(i)} = z_{(1)} \, .. \, z_{(i)} \bullet N^{(i)} \bullet z^{*}_{(1)} \, .. \, z^{*}_{(i-1)} I$.

If it is possible to choose the $x_{(i)}$ so that $x_{(i)} \bullet x^{*}_{(j)} = \delta_{ij}$, then it can be shown that the field $\mathcal{S}_1$ has the property that $\mu_1 \mu_1^{*} + .. + \mu_m \mu_m^{*} = 0$ implies $\mu_1 = .. = \mu_m = 0$ for each $\mu_1, .., \mu_m \in \mathcal{S}_1$.

A 2.8 Partial units in a regular system.

In this subsection and in A 2.9 the tensor system $\mathcal{S}$ will be assumed to be regular.

A 2.8a Definition. Let $x_{(1)\alpha}$, .., $x_{(\lambda)\alpha}$, $y^{\alpha}_{(1)}$, .., $y^{\alpha}_{(\lambda)}$ be as in A 2.7c and put

$$u_{(\alpha)}{}^{\alpha}_{\alpha} = y_{(1)} \, x_{(1)} + .. + y_{(\lambda)} \, x_{(\lambda)} .$$

Such an element $u_{(\alpha)}$ will be called a partial α-unit of $\mathcal{S}$.

A 2.8b Lemma. If p^{α}, $q_{\alpha} \in \mathcal{S}$, then $u_{(\alpha)} \bullet p = p$ and $q \bullet u_{(\alpha)} = q$.

Proof. Since the $y^{\alpha}_{(i)}$ span $\mathcal{S}_{\alpha *}$, $p = \lambda_1 \, y_{(1)} + .. + \lambda_\lambda \, y_{(\lambda)}$ for some λ_1 , .., $\lambda_\lambda \in \mathcal{S}_1$, whence $x_{(i)} \bullet p = \lambda_i$. Thus $u_{(\alpha)} \bullet p =$

$= \{ y_{(1)} \, x_{(1)} + .. + y_{(\lambda)} \, x_{(\lambda)} \} \bullet p = y_{(1)} \lambda_1 + .. + y_{(\lambda)} \lambda_\lambda = p.$

Similarly $q \bullet u_{(\alpha)} = q$.

A 2.8c Lemma. If $a^{\alpha}_{\beta} \in \mathcal{S}$, then $u_{(\alpha)} \bullet a = a$ and $a \bullet u_{(\beta)} = a$ are each necessary and sufficient conditions for a to be of the form

CHAPTER 6

$$a^{\alpha}_{\beta} = \sum_{i} p^{\alpha}_{(i)}\ q_{(i)\beta} \text{ with } p^{\alpha}_{(i)},\ q_{(i)\beta} \in \mathfrak{J} .$$

Proof. That the conditions are necessary follows from A 2.8b. Conversely, if $u_{(\alpha)} \cdot a = a$, then $a = \sum_{i} y_{(i)} (x_{(i)} \cdot a)$ which is of the required form. The proof for $a \cdot u_{(\beta)} = a$ is similar.

A 2.8d Lemma. For each $\alpha \in \mathfrak{S}$, $u_{(\alpha)} \cdot u_{(\alpha)} = u_{(\alpha)}$.

Proof. This follows from A 2.8c.

A 2.8e Lemma. For each $\alpha \in \mathfrak{S}$, the partial α-unit $u_{(\alpha)}$ is unique.

Proof. Suppose $u_{(\alpha)}$ and $u'_{(\alpha)}$ both satisfy A 2.8c. Then if $a^{\alpha}_{\alpha} \in \mathfrak{J}$, $a \cdot u_{(\alpha)} = a$ if and only if $u'_{(\alpha)} \cdot a = a$. Thus, if $u'_{(\alpha)}$ and $u_{(\alpha)}$ both satisfy A 2.8d we have, in particular, $u'_{(\alpha)} \cdot u_{(\alpha)} = u'_{(\alpha)}$ and $u'_{(\alpha)} \cdot u_{(\alpha)} = u_{(\alpha)}$.

A 2.8f Lemma. Let $a^{\alpha}_{\beta} \in \mathfrak{J}$; then $u_{(\alpha)} \cdot a = a \cdot u_{(\beta)}$.

Proof. By A 2.8c, $u_{(\alpha)} \cdot a = u_{(\alpha)} \cdot a \cdot u_{(\beta)}$, since $u_{(\alpha)} \cdot a = \sum_{i} y_{(i)} (x_{(i)} \cdot a)$. Similarly $a \cdot u_{(\beta)} = u_{(\alpha)} \cdot a \cdot u_{(\beta)}$.

A 2.8g Lemma. For each $\alpha \in \mathfrak{S}$, the dimension of the vector space $\mathfrak{J}_{\alpha}$ is equal to $\langle u_{(\alpha)} \rangle$.

Proof. This follows from A 2.8a and A 2.7c.

A 2.9 Theorem. Let $a^{\alpha}_{\beta} \in \mathfrak{J}$; then the number of linearly independent elements of the form $x \cdot a$, where $x_{\alpha} \in \mathfrak{J}_{\alpha}$, is equal to the rank of $u_{(\alpha)} \cdot a$.

Proof. Let s be the number of linearly independent elements of the form $x \cdot a$ and let $x_{(1)\alpha}, \ldots, x_{(s)\alpha}, y^{\alpha}_{(1)}, \ldots, y^{\alpha}_{(r)}$ be as in A 2.7c. Now, since $x_{(1)}, \ldots, x_{(r)}$ span $\mathfrak{J}_{\alpha}$, exactly s of the tensors

$x_{(1)} \cdot a, \ldots, x_{(\Delta)} \cdot a$ will be linearly independent. Without loss of generality, we can suppose these are $x_{(1)} \cdot a, \ldots, x_{(\Delta)} \cdot a$. Thus, by A 2.7,

$$0 \neq \{(x_{(1)} \cdot a)(x_{(2)} \cdot a) \ldots (x_{(\Delta)} \cdot a)\} \cdot N^{(\Delta)}$$

$$= \{(x_{(1)} \cdot u_{(\alpha)} \cdot a) \ldots (x_{(\Delta)} \cdot u_{(\alpha)} \cdot a)\} \cdot N^{(\Delta)}$$

$$= x_{(1)} \ldots x_{(\Delta)} \cdot (u_{(\alpha)} \cdot a)^{\Delta} \cdot N^{(\Delta)}$$

so that $(u_{(\alpha)} \cdot a)^{\Delta} \cdot N^{(\Delta)} \neq 0$. Also by A 2.7,

$$(u_{(\alpha)} \cdot a)^{\Delta+1} \cdot N^{(\Delta+1)} = \sum_{i_1, \ldots, i_{\Delta+1}=1}^{n} (y_{(i_1)} x_{(i_1)} \cdot a) \ldots (y_{(i_{\Delta+1})} x_{(i_{\Delta+1})} \cdot a) \cdot N^{(\Delta+1)}$$

$$= \sum_{i_1, \ldots, i_{\Delta+1}=1}^{n} y_{(i_1)} \ldots y_{(i_{\Delta+1})} \{(x_{(i_1)} \cdot a) \ldots (x_{(i_{\Delta+1})} \cdot a) \cdot N^{(\Delta+1)} = 0.$$

A 2.9a Corollary. If $a \begin{matrix} \alpha_1 \\ \alpha_2 \end{matrix} \in \mathfrak{J}$ and $a \cdot a = a$, then the number of linearly independent elements of the form $x \cdot a$, where $x_{\alpha} \in \mathfrak{J}_{\alpha}$, is equal to $\langle u_{(\alpha)} \cdot a \rangle$.

Proof. By A 2.8f and A 2.8g, $(u_{(\alpha)} \cdot a) \cdot (u_{(\alpha)} \cdot a) = u_{(\alpha)} \cdot u_{(\alpha)} \cdot a \cdot a$

$$= u_{(\alpha)} \cdot a .$$

Thus, the result follows from theorem A 2.9 by A 2.8c.

A 2.9b Corollary. Let $\mathfrak{J}$ be normal and let $a \begin{matrix} \alpha \\ \beta \end{matrix} \in \mathfrak{J}$; then the number of linearly independent elements of the form $x \cdot a$, where $x_{\alpha} \in \mathfrak{J}_{\alpha}$, is equal to $\langle u_{(\alpha)} \cdot a \cdot a^{\dagger} \rangle = \langle u_{(\beta)} \cdot a^{\dagger} \cdot a \rangle$.

Proof. Since $a \cdot a^{\dagger} \cdot a = a$, the number of linearly independent elements of the form $x \cdot a \cdot a^{\dagger}$ is the same as that of elements of the form $x \cdot a$.

The result follows from A 2.9a since $\langle u_{(\alpha)} \cdot a \cdot a^{\dagger} \rangle = \langle a \cdot u_{(\beta)} \cdot a^{\dagger} \rangle$ $= \langle u_{(\beta)} \cdot a^{\dagger} \cdot a \rangle$ by A 2.8f and A 1.3o.

A 2.9c Corollary Let $a_{\beta}^{\alpha} \in \mathcal{J}$ and $a \cdot a = a$; then the number of linearly independent elements x_{α} of $\mathcal{J}_{\alpha}$ which satisfy $x \cdot a = 0$ is equal to $\langle u_{(\alpha)} \rangle - \langle a \rangle$.

Proof. This follows from A 2.8g and A 2.9a.

A 2.9d Corollary. If $\mathcal{J}$ is normal and a_{β}^{α} , $c_{\gamma}^{\alpha} \in \mathcal{J}$, the number of linearly independent solutions in $\mathcal{J}$ of the equation $a \cdot x = c$ is $\langle u_{(\beta)} \rangle - \langle u_{(\beta)} \cdot a^{\dagger} \cdot a \rangle$, provided that the equation has a solution.

Proof. If the equation has a solution, the number of linearly independent solutions is the same as that for $a \cdot x = 0$. Hence the result follows from A 2.9c.

A 2.10 Group tensor systems.

In this subsection I shall show how any finite group, together with a field $\mathfrak{A}$, may be embedded in a tensor system analogous to a group ring over $\mathfrak{A}$. It is possible to do this for an arbitrary field $\mathfrak{A}$, but in order to produce a normal tensor system I shall place a restriction on the field $\mathfrak{A}$. This will enable all the preceding results concerning linear equations to be applied in such a tensor system and will also simplify considerations of identities in the system.

A 2.10a Definition. A field $\mathfrak{R}$ will be called formally complex if it possesses an involution $\lambda \to \lambda^*$ (i.e. $(\lambda)^* = \lambda$, $(\lambda\mu)^* = \lambda^*\mu^*$, $(\lambda + \mu)^* = \lambda^* + \mu^*$) with the property that for any set $\lambda_1, .., \lambda_\nu \in \mathfrak{R}$

$$\lambda_1\lambda_1^* + .. + \lambda_\nu\lambda_\nu^* = 0 \text{ implies } \lambda_1 = .. = \lambda_\nu = 0$$

(see A 2.7e). This is analogous to the idea of a formally real field (see van der Waerden (22) p.225).

With this definition the fields of complex, real, algebraic, or rational numbers are all formally complex. Any formally complex field is clearly without characteristic since $1^* = 1$.

A 2.10b Let $\mathfrak{G}$ be a finite group of order g and let $\mathfrak{R}$ be a formally complex field. Let the semigroup $\mathfrak{S}$ be semigenerated by the single index type α with $\nu(\alpha) = g$. The group tensor system $\mathcal{T} = \mathcal{T}(\mathfrak{G}, \mathfrak{R})$ of $\mathfrak{G}$ over $\mathfrak{R}$ is a tensor system over $\mathfrak{S}$ defined by the following properties A 2.10c, .., A 2.10f.

A 2.10c The field $\mathcal{T}_1$ is isomorphic with $\mathfrak{R}$ and will be identified with $\mathfrak{R}$. The set $\mathcal{T}_{\alpha\alpha^*}$ is isomorphic with the group algebra of $\mathfrak{G}$ over $\mathfrak{R}$ and it will be identified with this group algebra, the product of two elements of the group algebra being written as the contracted product of elements of $\mathcal{T}_{\alpha\alpha^*}$. Thus, if $a_{(1)}, .., a_{(g)}$ are the elements of $\mathfrak{G}$, the elements of $\mathcal{T}_{\alpha\alpha^*}$ are of the form $\sum_{i=1}^{g} \lambda_i a_{(i)\alpha}^{\alpha}$ with $\lambda_i \in \mathfrak{R}$. The product of $a_{(i)}$ with $a_{(j)}$ is $a_{(i)} \cdot a_{(j)} = a_{(i)\bar{\alpha}}^{\alpha} a_{(j)\alpha}^{\bar{\alpha}} = a_{(k)\alpha}^{\alpha}$, say.

A 2.10d The trace $\langle a_{(i)} \rangle = a_{(i)} \frac{\bar{\alpha}}{\alpha}$ of the group element $a_{(i)}$ is defined to be zero unless $a_{(i)} = I_{(\alpha)}$, the identity of $\mathfrak{G}$. In this case we have $\langle I \rangle = \nu(\alpha) = g$. Thus this trace operation is just the usual trace operation for a group algebra (see Littlewood (11) p.47 and A 1.3k).

A 2.10e The elements $a_{(1)}, \ldots, a_{(g)}$ of $\mathfrak{G}$ and the elements of $\mathfrak{R}$ generate $\mathfrak{T}$. Thus if $r \neq s$, the set $\mathfrak{T}_{\alpha^r \alpha^{*s}}$ consists of a zero only, while the elements of $\mathfrak{T}_{\alpha^r \alpha^{*r}}$ are of the form

$$\sum_{i_1, \ldots, i_r, k} \lambda_{(i_1, \ldots, i_r; k)} a_{(i_1)} \overset{\alpha_1}{\alpha_{k(1)}} \cdots a_{(i_r)} \overset{\alpha_r}{\alpha_{k(r)}}$$

the k's being permutations of 1, .., r and where the λ's are elements of $\mathfrak{R}$. The sum, product, or transpose of any two such expressions is clearly again such an expression. Also, in virtue of the definitions given in A 2.10c and A 2.10d, any contraction of such an expression gives rise to another such expression.

A 2.10f The system $\mathfrak{T}$ is normal, the involution on $\mathfrak{T}_1$ being the involution $\lambda \to \lambda^*$ of $\mathfrak{R}$.

The conjugate of an element of $\mathfrak{G}$ is defined as its inverse. This determines the involution completely.

A 2.10g Identities in $\mathfrak{T}(\mathfrak{G}, \mathfrak{R})$. As is the case with elements of $\mathfrak{J}$, there are elements of $\mathfrak{T}$ which are zero but which do not vanish formally. This is necessitated by the fact that $\mathfrak{T}$ is required to be normal. Similarly to the case of elements of $\mathfrak{J}$, an algorithm (see A 1.7c) for determining

these identities may be given. Thus if

$$c_{\alpha^{\wedge}}^{\alpha^{\wedge}} = \sum \lambda_{(i_1,\ldots,i_\lambda;k)}\ a_{(i_1)\,\alpha_{k(1)}}^{\alpha_1} \cdots a_{(i_\lambda)\,\alpha_{k(\lambda)}}^{\alpha_1}$$

it can be decided whether or not $c = 0$ by forming $\langle c \cdot c^* \rangle$, which is a polynomial in g and the λ's and λ^*'s, and which is zero if, and only if, $c = 0$.

This gives us a rule for deciding whether or not two expressions representing elements of $\mathcal{T}$ are equal, that is, whether their difference is zero. This rule is defined uniquely by the definitions A 2.10c, .., A 2.10f, so that the system $\mathcal{T}(\mathfrak{G}, \mathfrak{R})$ is defined uniquely provided that these definitions do not lead to an inconsistency.

A 2.10h Consistency of $\mathcal{T}(\mathfrak{G}, \mathfrak{R})$. The consistency of a system subject to A 2.10b, .., A 2.10e may be demonstrated by giving a representation of $\mathcal{T}$ in terms of a representation of $\mathfrak{G}$ by regular group matrices (see Littlewood (11) p.53). Thus the group element $a_{(i)}$ is represented by the matrix whose $(r, s)^{**}$ component $b_{(i)\,s}^{\ \ r}$ is given by

$$b_{(i)\,\Delta}^{\ \ \wedge} = \begin{cases} 1 \text{ if } a_{(\lambda)} = a_{(i)} \cdot a_{(\Delta)} \\ 0 \text{ otherwise.} \end{cases}$$

Let $\mathcal{S}$ be the tensor system generated by these matrices and the field $\mathfrak{R}$, each matrix $(b_{(i)\,\Delta}^{\ \ \wedge})$ being interpreted as an element $b_{(i)\,\alpha}^{\ \ \alpha}$ of $\mathcal{S}_{\alpha\alpha^*}$. We have $b_{(i)} \cdot b_{(j)} = b_{(k)}$ where $a_{(i)} \cdot a_{(j)} = a_{(k)}$ since $\sum_t b_{(i)\,t}^{\ \ \wedge}\ b_{(j)\,\Delta}^{\ \ t} = b_{(k)\,\Delta}^{\ \ \wedge}$. Also $\langle b_{(i)} \rangle = g$ or 0 according as $a_{(i)}$ is

the identity of $\mathfrak{G}$ or not. The matrices $b_{(i)}$ are linearly independent, so that $\mathfrak{J}_{\alpha\kappa^*}$, consisting of elements of the form $\sum_i \lambda_i b_{(i)}$, is isomorphic with the group algebra of $\mathfrak{G}$ over $\mathfrak{K}$. Thus $\mathfrak{J}$ possesses each of the properties A 2.10c, .., A 2.10e required of $\mathfrak{T}$. Now suppose $\mathfrak{J}$ possesses an involution which is defined on the elements of $\mathfrak{K}$ by $\lambda \longrightarrow \lambda^*$ and on $b_{(1)}$, .., $b_{(g)}$ by $(b_{(i)}^*)_{\alpha}^{\lambda} = b_{(i)\,\lambda}^{\alpha}$. This extends to any element $c_{\alpha^\rho}^{\alpha^\rho} = \sum \lambda_{(i_1,\ldots,i_\lambda;k)} a_{(i_1)\,\alpha_{k(1)}}^{\alpha_1} \ldots a_{(i_\lambda)\,\alpha_{k(\lambda)}}^{\alpha_\lambda}$ of $\mathfrak{J}$ so that, if c is given by the set of elements $c_{\alpha_1 \ldots \alpha_\rho}^{\lambda_1 \ldots \lambda_\rho}$ of $\mathfrak{K}$, then c^* is given by the corresponding set $c_{\lambda_1 \ldots \lambda_\rho}^{\alpha_1 \ldots \alpha_\rho *}$. The scalar $\langle c \cdot c^* \rangle$ is given by $\sum c_{\alpha_1 \ldots \alpha_\rho}^{\lambda_1 \ldots \lambda_\rho} c_{\alpha_1 \ldots \alpha_\rho}^{\lambda_1 \ldots \lambda_\rho *}$ and hence by A 2.10a, $\langle c \cdot c^* \rangle = 0$ only if $c = 0$. It follows that $\mathfrak{J}$ is normal. Thus, since $b_{(i)}^*$ is the inverse of $b_{(i)}$, $\mathfrak{J}$ possesses the property A 2.10f also required of $\mathfrak{T}$. This proves the consistency of $\mathfrak{T}(\mathfrak{G}, \mathfrak{K})$.

A 2.10i Since $\mathfrak{T}$ is a normal system, it follows that a generalized inverse (see A 2.2) exists for each element of a group algebra over $\mathfrak{K}$. In particular, this applies to elements of $\mathfrak{J}$ if $\mathfrak{J}_{(0)} = \mathfrak{K}$. The recurrence relation given in A 2.4 for computing this generalized inverse also applies in a group algebra and hence any linear equation $c_{(1)} \cdot x \cdot d_{(1)} + \ldots + c_{(\lambda)} \cdot x \cdot d_{(\lambda)} = f$ may be solved by means of theorem A 2.5 (see A 2.5b).

A 2.11 Theorem. The partial $\alpha\alpha^*$-unit $u_{(\alpha\alpha^*)}$ of $\mathcal{T}(\mathfrak{G},\mathfrak{K})$ is given by

$$u_{(\alpha\alpha^*)}{}^{\alpha_1\alpha_1^*}_{\alpha_2\alpha_2^*} = \frac{1}{g}\sum_{i=1}^{g} a_{(i)}{}^{\alpha_1}_{\alpha_1}\ a^*_{(i)}{}^{\alpha_2}_{\alpha_2} .$$

Proof. The elements $x_{(i)\,\alpha\alpha^*} = \frac{1}{g}\, a_{(i)}{}^{\alpha}_{\alpha}$ and $y_{(i)}{}^{\alpha\alpha^*} = a^*_{(i)}{}^{\alpha}_{\alpha}$ for i = 1, .., g satisfy the conditions of A 2.7c since

$\frac{1}{g}\, a_{(i)}{}^{\bar\alpha_1}_{\bar\alpha_2}\ a^*_{(j)}{}^{\bar\alpha_2}_{\bar\alpha_1} = \frac{1}{g}\langle a_{(i)} \cdot a^*_{(j)}\rangle = \delta_{ij}$ by A 2.10c, A 2.10d and A 2.10f.

The result follows from A 2.8a.

A 2.11a Corollary. Let c^{α}_{α}, $d^{\alpha}_{\alpha} \in \mathcal{T}_{\alpha\alpha^*}$. Then the number of linearly independent expressions of the form c•x•d, where $x^{\alpha}_{\alpha} \in \mathcal{T}_{\alpha\alpha^*}$ is given by $\frac{1}{g}\sum_{i=1}^{g}\langle c^{\dagger}\cdot c\cdot a_{(i)}\cdot d\cdot d^{\dagger}\cdot a^*_{(i)}\rangle$.

Proof. By A 2.9b and A 2.2m, the number of linearly independent expressions $(c^{\alpha_1}_{\bar\alpha_1}\ d^{\bar\alpha_2}_{\alpha_2})x^{\bar\alpha_1}_{\bar\alpha_2}$ is given by $u_{(\alpha\alpha^*)}{}^{\bar\alpha_3\bar\alpha_4^*}_{\bar\alpha_1\bar\alpha_2^*}\ c^{\dagger}{}^{\bar\alpha_1}_{\bar\alpha_3}\ d^{\dagger}{}^{\bar\alpha_4}_{\bar\alpha_2}\ c^{\bar\alpha_3}_{\bar\alpha_5}\ d^{\bar\alpha_5}_{\bar\alpha_4}$

The result follows from A 2.11.

A 2.11b Corollary. Let e^{α}_{α} be an idempotent element of $\mathcal{T}_{\alpha\alpha^*}$. Then the number of linearly independent solutions of e•x = 0 with $x^{\alpha}_{\alpha} \in \mathcal{T}_{\alpha\alpha^*}$ is given by g − ⟨e⟩.

Proof. This follows from A 2.11a with c = e and d = I since $\langle e^{\dagger}\cdot e\cdot a_{(i)}\cdot a^*_{(i)}\rangle = \langle e^{\dagger}\cdot e\cdot e\rangle = \langle e\cdot e^{\dagger}\cdot e\rangle = \langle e\rangle$. Alternatively, it follows directly from A 2.11 by A 2.9a .

The result of A 2.11b is due to Rutherford ((20) p.87) if $\mathfrak{G}$ is a symmetric group.

A 2.11c By means of theorem A 2.11 and A 2.9d the number of linearly independent solutions of any linear equation in elements of a group algebra (over a formally complex field) may be obtained and this applies in particular to elements of $\mathfrak{J}$ (if $\mathfrak{J}_{(0)}$ is formally complex).

A 3 Complete tensor systems.

In this chapter I show how a simple algebraic restriction may be placed on a tensor system, to give a system isomorphic with a complete system of "ordinary" tensors. Certain properties of such systems are discussed and simultaneous linear equations are treated.

A 3.1 Tensors representable by sets of components.

A 3.1a Definition. A complete tensor system will be defined as a proper tensor system $\mathfrak{C}$ having the property that for each non-zero $a^{\alpha}_{\beta} \in \mathfrak{C}$, there exists an element $x^{\beta} \in \mathfrak{C}$ such that $a^{\alpha}_{\bar{\beta}}\, x^{\bar{\beta}} \neq 0$.

It is clear that any complete system is weakly regular (see A 1.6t) since we can take $\alpha = 1$. Hence by A 2.1b, any complete system is regular. Also (see A 2.8a), it contains a partial unit $u_{(\alpha)}$ for each index type α.

A 3.1b Lemma. Let $\mathfrak{C}$ be a complete system. Then the partial units are given by $u_{(\alpha)} = I_{(\alpha)}$ for each α.

Proof. Suppose $I - u_{(\alpha)} \neq 0$. Then, by A 3.1a, there exists $x^{\alpha} \in \mathfrak{C}$

such that $(I - u_{(\alpha)}) \cdot x \neq 0$. This contradicts A 2.8b.

A 3.1c Lemma. If $\mathfrak{C}$ is complete, the dimension of the space $\mathfrak{C}_\alpha$ is $\nu(\alpha)$.

Proof. This follows from A 2.8g and A 3.1b (see A 1.4e).

A 3.1d Definition. Let $\mathfrak{C}$ be a complete system over $\mathfrak{G}$ and let $\rho, \sigma, \ldots$ semigenerate $\mathfrak{G}$. An ordered set $\mathfrak{F}$:

$$x_{(\rho,1)\rho}, \ldots, x_{(\rho,\nu(\rho))\rho} ; \; y^{\rho}_{(\rho,1)}, \ldots, y^{\rho}_{(\rho,\nu(\rho))} ; \; x_{(\sigma,1)\sigma}, \ldots, x_{(\sigma,\nu(\rho))\sigma}, \ldots$$

of elements of $\mathfrak{C}$ will be called a frame for $\mathfrak{C}$ if

$$\sum_{i=1}^{\nu(\rho)} y_{(\rho,i)} x_{(\rho,i)} = I_{(\rho)}, \quad \sum_{i=1}^{\nu(\sigma)} y_{(\sigma,i)} x_{(\sigma,i)} = I_{(\sigma)}, \ldots \qquad \text{A 3.1e}$$

and

$$x_{(\rho,i)} \cdot y_{(\rho,j)} = \delta_{ij}, \; x_{(\sigma,i)} \cdot y_{(\sigma,j)} = \delta_{ij}, \ldots \qquad \text{A 3.1f}$$

Any complete system possesses at least one frame, by A 2.7c, A 2.8a, A 3.1b and A 3.1c.

A 3.1g Definition. Let $\mathfrak{C}$ and $\mathfrak{F}$ be as in A 3.1d and let $a^{\rho_1 \ldots \rho_\lambda \; \sigma_1 \ldots \sigma_\mu \ldots}_{\rho_1 \ldots \rho_{\lambda'} \; \sigma_1 \ldots \sigma_{\mu'} \ldots} \in \mathfrak{C}$. The scalar $a^{i_1 \ldots i_\lambda \; j_1 \ldots j_\mu \ldots}_{i'_1 \ldots i'_{\lambda'} \; j'_1 \ldots j'_{\mu'} \ldots}$

$$= \{x_{(\rho,i_1)} \cdots x_{(\rho,i_\lambda)} x_{(\sigma,j_1)} \cdots x_{(\sigma,j_\mu)} \ldots\} \cdot a \cdot \{y_{(\rho,i'_1)} \ldots y_{(\rho,i'_{\lambda'})} \; y_{(\sigma,j'_1)} \ldots\}$$

will be called the $(i_1, \ldots, i_\lambda ; i'_1, \ldots, i'_{\lambda'} ; j_1, \ldots, j_\mu ; j'_1, \ldots, j'_{\mu'} ; \ldots)$-component of a in the frame $\mathfrak{F}$.

A 3.1h Lemma. Each element of a complete system $\mathfrak{C}$ is completely determined by its components in any given frame $\mathfrak{F}$.

Proof. Let $a^{\rho_1 \cdots \rho_\lambda \ \sigma_1 \cdots \sigma_\lambda \cdots}_{\rho_1 \cdots \rho_{\lambda'} \ \sigma_1 \cdots \sigma_{\lambda'} \cdots} \in \mathcal{C}$ and let $\mathcal{F}$ be as in A 3.1d. Then, by A 3.1b and A 2.8a, we have

$$a = \{(u_{(\rho)})^\lambda (u_{(\sigma)})^\lambda \cdots\} \cdot a \cdot \{(u_{(\rho)})^{\lambda'} (u_{(\sigma)})^{\lambda'} \cdots\}$$

$$= \sum y_{(\rho,i_1)} \cdots y_{(\rho,i_\lambda)} \, y_{(\sigma,j_1)} \cdots \{x_{(\rho,i_1)} \cdots x_{(\rho,i_\lambda)} x_{(\sigma,j_1)} \cdots \cdot a \cdot y_{(\rho,i'_1)} \cdots\} \, x_{(\rho,i'_1)} \cdots .$$

Hence

$$a = \sum a^{i_1 \cdots i_\lambda \, j_1 \cdots j_\lambda \cdots}_{i'_1 \cdots i'_{\lambda'} \, j'_1 \cdots j'_{\lambda'} \cdots} y_{(\rho,i_1)} \cdots y_{(\rho,i_\lambda)} \, y_{(\sigma,j_1)} \cdots x_{(\rho,i'_1)} \cdots x_{(\rho,i'_{\lambda'})} \, x_{(\sigma,j'_1)} \cdots \quad \text{A 3.1i}$$

the sums being taken over all sets of i's and j's, etc. for which the i's are taken from $1, \ldots, \nu(\rho)$, the j's from $1, \ldots, \nu(\sigma)$, etc. Thus, a is determined by its components and the elements of $\mathcal{F}$.

Note also that any expression of the form A 3.1i, where the $a^{i \cdots}_{i' \cdots}$'s are elements of $\mathcal{C}_1$, is an element of $\mathcal{C}$.

A 3.1j Corollary. The only elements of $\mathcal{C}$ all of whose components in $\mathcal{F}$ are zero, are the zero elements of $\mathcal{C}$.

A 3.1k By A 3.1h and A 3.1i we see that the elements of $\mathcal{C}$ can be regarded as an "ordinary" system of tensors. That is, a system in which each element of $\mathcal{C}$ has a representation as a set of components in some frame and where each component may be chosen arbitrarily from a field $\mathfrak{R}$ ($= \mathcal{C}_1$). The operations of transposition addition and outer product are represented by the corresponding operations on the sets of components. Contraction is represented by a summation over repeated indices (this follows from A 3.1i by A 3.1f).

A 3.11 Conversely, the consistency of complete tensor systems is demonstrated by the existence of systems which are represented in this way by sets of components. Such systems clearly give rise to complete tensor systems.

A 3.2 Theorem. Any complete tensor system $\mathfrak{C}$ over a semigroup $\mathfrak{S}$ is uniquely determined, up to isomorphism, by the field $\mathfrak{C}_1$ and the unordered set of integers $\nu(\rho)$, $\nu(\sigma)$, .., where ρ, σ, .. semigenerate $\mathfrak{S}$.

Proof. Let $\mathfrak{C}$ be a complete system over $\mathfrak{S}$ and let $\mathfrak{C}'$ be a complete system over $\mathfrak{S}'$, where $\mathfrak{C}_1 = \mathfrak{C}_1'$ and where the dimensions of the semigenerators ρ, σ, .. of $\mathfrak{S}$ are the dimensions of the semigenerators ρ', σ', .. of $\mathfrak{S}'$ in some order. Without loss of generality we have $\nu(\rho) = \nu(\rho')$, $\nu(\sigma) = \nu(\sigma')$, ... Let $\mathfrak{F}$ be a frame for $\mathfrak{C}$ and $\mathfrak{F}'$ a frame for $\mathfrak{C}'$. Let a be any element of $\mathfrak{C}$. Then the corresponding element a′ of $\mathfrak{C}'$ may be chosen which has the same components in $\mathfrak{F}'$ as a has in $\mathfrak{F}$. This correspondence $a \leftrightarrow a'$ establishes an isomorphism between $\mathfrak{C}$ and $\mathfrak{C}'$ since the tensor operations are completely determined in each of $\mathfrak{C}$ and $\mathfrak{C}'$, by A 3.1f and A 3.11, from the sets of components.

A 3.2a The complete tensor system over $\mathfrak{S}$, for which $\mathfrak{C}_1$ is a given field $\mathfrak{R}$ without characteristic, will be denoted by $\mathfrak{C}(\mathfrak{S}, \mathfrak{R})$. This system is essentially unique by A 3.2, since the integers $\nu(\rho)$, .. are to be

associated with the semigroup $\mathfrak{S}$ (see A 1.7d).

A 3.2b Henceforth, all tensor systems considered will be supposed to be subsystems of some $\mathcal{C}(\mathfrak{S}, \mathfrak{K})$.

A 3.2c Normal complete systems. Let $\mathfrak{K}$ be a formally complex field (see A 2.10a) and let $\mathcal{F}$ be a frame for the system $\mathcal{C} = \mathcal{C}(\mathfrak{S}, \mathfrak{K})$. Write $x^{*}_{(\rho,i)} = y_{(\rho,i)}$, $x^{*}_{(\sigma,i)} = y_{(\sigma,i)}$, etc., where $x_{(\rho,1)}$, .., $y_{(\rho,1)}$, .., $x_{(\sigma,1)}$, ... are the elements of $\mathcal{F}$. Then the involution $\lambda \longrightarrow \lambda^{*}$ of $\mathfrak{K}$ can be extended to the whole of $\mathcal{C}$ by defining the components $a^{*}_{\cdots}$ of a^{*} by $a^{* i_1' \cdots i_\lambda' j_1' \cdots j_\alpha' \cdots}_{i_1 \cdots i_\lambda \, j_1 \cdots j_\alpha \cdots} = (a^{i_1 \cdots i_\lambda \, j_1 \cdots j_\alpha \cdots}_{i_1' \cdots i_\lambda' \, j_1' \cdots j_\alpha' \cdots})^{*}$, $a^{\cdots}_{\cdots}$ being the components in $\mathcal{F}$ of the element a.

A 3.2d If $\mathcal{C}$ is normal and $\mathcal{F}$ is a frame for $\mathcal{C}$ for which $x^{*}_{(\rho,i)} = y_{(\rho,i)}$, etc., then $\mathcal{F}$ will be called a normal frame for $\mathcal{C}$. A subsystem $\mathcal{S}$ of $\mathcal{C}$ will be called a normal subsystem of $\mathcal{C}$ if it is normal with the same involution as $\mathcal{C}$. Thus $\mathcal{J}$ and $\mathcal{R}$ are normal subsystems of (see A 1.6p).

A 3.2e Remark. It is possible to show that any complex normal system (see A 1.6s), which possesses α-units for each index type α, is necessarily a proper system, i.e. the $\nu(\alpha)$'s are necessarily integers. It follows, therefore, that for complex normal complete systems, the restriction that $\mathcal{C}$ be proper is not necessary, provided that $\mathcal{C}$ possesses α-units for each index type.

A 3.3 Automorphisms of $\mathcal{C}(\mathfrak{S}, \mathfrak{R})$

An automorphism Φ of a tensor system is a 1-1 mapping of the system onto itself which preserves the tensor operations. I shall be concerned, here, with automorphisms of $\mathcal{C} = \mathcal{C}(\mathfrak{S}, \mathfrak{R})$. An automorphism Φ, of $\mathcal{C}$, has the following properties.

A 3.3a For each $\alpha \in \mathfrak{S}$, Φ is a 1-1 mapping of $\mathcal{C}_\alpha$ onto itself.

A 3.3b If a' is a transpose of a, then $\Phi(a')$ is the corresponding transpose of $\Phi(a)$.

A 3.3c $\Phi(a + b) = \Phi(a) + \Phi(b)$.

A 3.3d $\Phi(ab) = \Phi(a)\ \Phi(b)$.

A 3.3e If b is a contraction of a, then $\Phi(b)$ is the corresponding contraction of $\Phi(a)$. Thus, we have $\Phi(a \cdot b) = \Phi(a) \cdot \Phi(b)$ and $\Phi(\langle a \rangle) = \langle \Phi(a) \rangle$.

A 3.3f The automorphism Φ will be called a linear transformation if it satisfies A 3.3a, .., A 3.3e and also, $\Phi(\lambda) = \lambda$ for all $\lambda \in \mathcal{C}_1$.

A 3.3g Lemma. The mapping $\mathcal{C}_1 \to \Phi(\mathcal{C}_1)$ is an automorphism of $\mathcal{C}_1$.

Proof. This is immediate from A 3.3a, A 3.3c and A 3.3e.

A 3.3h Lemma. For all $\alpha \in \mathfrak{S}$, $\Phi(I_{(\alpha)}) = I_{(\alpha)}$.

Proof. This follows from A 3.3e and the uniqueness of $I_{(\alpha)}$ (A 1.4b).

A 3.3i Lemma. $\Phi(\mathcal{F}) = \mathcal{F}$.

Proof. This follows from A 3.3h since $\mathcal{F}$ is generated by $\mathcal{C}_1$ and the units of $\mathcal{C}$.

A 3.3j Lemma. If $A \in \mathcal{R}$, then $\Phi(A) = A$.

Proof. The rational elements $\mathcal{R}_1$ of $\mathcal{C}_1$ are invariant under Φ, by A 3.3g. Now $\mathcal{R}$ is generated by ($\mathcal{R}_1$ and) the units of $\mathcal{C}$, so that its elements are also invariant under Φ. Similarly, there is the following result.

A 3.3k Lemma. If Φ is a linear transformation, then for each element $A \in \mathcal{F}$, $\Phi(A) = A$.

A 3.3l Lemma. Given any automorphism of $\mathcal{C}_1$ and any two frames $\mathcal{F}$ and $\mathcal{F}'$ of $\mathcal{C}$, there exists a unique automorphism Φ of $\mathcal{C}$ which induces the given automorphism on $\mathcal{C}_1$ and which sends $\mathcal{F}$ into $\mathcal{F}'$.

Proof. By A 3.1h, any element of $\mathcal{C}$ is completely determined by its components in $\mathcal{F}$ and by the elements of $\mathcal{F}$. Thus, if Φ is defined to map the elements of $\mathcal{F}$ into the corresponding elements of $\mathcal{F}'$ and the elements of $\mathcal{C}_1$ according to the given automorphism, this determines Φ uniquely as an automorphism of $\mathcal{C}$. (The formal properties of the elements of $\mathcal{F}$ are the same as those of $\mathcal{F}'$.)

A 3.3m Lemma. Let $x_{(1)\rho}, \ldots, x_{(\lambda)\rho}$; $x'_{(1)\rho}, \ldots, x'_{(\lambda)\rho}$ be two sets of linearly independent elements of $\mathcal{C}_\rho$, ρ being a generator of $\mathfrak{G}$. Then there exists a linear transformation Φ of $\mathcal{C}$ such that $\Phi(x_{(i)}) = x'_{(i)}$

$(i = 1, .., r)$.

Proof. By A 2.7c, we can choose frames $\mathfrak{F}$ and $\mathfrak{F}'$ of $\mathfrak{C}$ such that $x_{(1)}, .., x_{(\lambda)} \in \mathfrak{F}$ and $x'_{(1)}, .., x'_{(\lambda)} \in \mathfrak{F}'$. Thus the result follows from A 3.3e with the identity automorphism of $\mathfrak{C}_i$.

A 3.3n If $\rho, \sigma, ..$ semigenerate $\mathfrak{G}$, we see that any linear transformation Φ of $\mathfrak{C}$ is defined by a set of linear transformations of the vector spaces $\mathfrak{C}_{\rho^*}, \mathfrak{C}_{\sigma^*}, ...$ Thus, if p^{ρ}_{ρ}, q^{σ}_{σ}, .. are the respective matrices of transformation of these spaces and $p^{(-1)\rho}_{\rho}$, $q^{(-1)\rho}_{\rho}$, .. are their inverses and if $a^{\rho\lambda\sigma\mu\ldots}_{\rho'\lambda'\sigma'\mu'\ldots} \in \mathfrak{C}$, then we have

$$\Phi(a) = \{(p)^{\lambda}(q)^{\mu}..\} \bullet a \bullet \{(p^{(-1)})^{\lambda'}(q^{(-1)})^{\mu'}..\}.$$

A 3.3o A linear transformation Φ is closely connected with the corresponding transformation of the components of the elements of $\mathfrak{C}$ when the frame $\mathfrak{F}$ of $\mathfrak{C}$ is replaced by its inverse image $\Phi^{-1}(\mathfrak{F})$. Thus the expression for $\Phi(a)$ given in A 3.3n gives the usual form for the transformation of tensor components under change of coordinate system.

A 3.3p Lemma. The elements of $\mathfrak{J}$ have the same components in any frame.

Proof. This follows from A 3.3k (see A 3.3o).

A 3.4 Tensor matrices.

Let $\mathfrak{C}$ be a normal complete system and let $a_{(1,1)\zeta}^{\alpha}$, .., $a_{(1,\lambda)\theta}^{\alpha}$; ..... ; $a_{(\mu,1)\zeta}^{\gamma}$, .., $a_{(\mu,\lambda)\theta}^{\gamma}$ be elements of $\mathfrak{C}$. The array

$$\mathcal{A} = \begin{pmatrix} a_{(1,1)} & \cdots & a_{(1,\lambda)} \\ \vdots & & \vdots \\ a_{(\alpha,1)} & \cdots & a_{(\alpha,\lambda)} \end{pmatrix}$$

will be called a tensor matrix over $\mathcal{C}$. (It is a matrix whose elements are tensors.)

Let $\mathcal{B} = \begin{pmatrix} b_{(1,1)} & \cdots & b_{(1,t)} \\ \vdots & & \vdots \\ b_{(u,1)} & \cdots & b_{(u,t)} \end{pmatrix}$. Then $\mathcal{A}$ and $\mathcal{B}$ will be conformable for addition if and only if $r = t$, $s = u$, and each sum $a_{(i,j)} + b_{(i,j)}$ exists. The sum of $\mathcal{A}$ and $\mathcal{B}$ is then given, as with ordinary matrices, by

$$\mathcal{A} + \mathcal{B} = \begin{pmatrix} a_{(1,1)} + b_{(1,1)} & \cdots & a_{(1,\lambda)} + b_{(1,\lambda)} \\ \vdots & & \vdots \\ a_{(\alpha,1)} + b_{(\alpha,1)} & \cdots & a_{(\alpha,\lambda)} + b_{(\alpha,\lambda)} \end{pmatrix}.$$

Similarly, $\mathcal{A}$ and $\mathcal{B}$ are conformable for multiplication (in that order) if and only if $r = u$ and the products $a_{(i,\lambda)} \cdot b_{(\lambda,j)}$ all exist. The product of $\mathcal{A}$ with $\mathcal{B}$ is then

$$\mathcal{A} \cdot \mathcal{B} = \begin{pmatrix} a_{(1,1)} \cdot b_{(1,1)} + \cdots + a_{(1,\lambda)} \cdot b_{(\lambda,1)} & \cdots & a_{(1,1)} \cdot b_{(1,t)} + \cdots + a_{(1,\lambda)} \cdot b_{(\lambda,t)} \\ \vdots & & \vdots \\ a_{(\alpha,1)} \cdot b_{(1,1)} + \cdots + a_{(\alpha,\lambda)} \cdot b_{(\lambda,1)} & \cdots & a_{(\alpha,1)} \cdot b_{(1,t)} + \cdots + a_{(\alpha,\lambda)} \cdot b_{(\lambda,t)} \end{pmatrix}.$$

That the product $\mathcal{A} \cdot \mathcal{B}$ exists, follows from the given restriction on the types of the elements of a tensor matrix.

The involution operation of $\mathcal{C}$ can also be applied to tensor matrices. We have

$$\mathcal{A}^* = \begin{pmatrix} a_{(1,1)} & \cdots & a_{(1,\lambda)} \\ \vdots & & \vdots \\ a_{(\alpha,1)} & \cdots & a_{(\alpha,\lambda)} \end{pmatrix}^* = \begin{pmatrix} a^*_{(1,1)} & \cdots & a^*_{(\alpha,1)} \\ \vdots & & \vdots \\ a^*_{(1,\lambda)} & \cdots & a^*_{(\alpha,\lambda)} \end{pmatrix}.$$

When the elements $a_{(i,j)}$ are all scalars, the tensor matrix becomes an ordinary matrix defined over $\mathfrak{C}_1$. The matrix $\mathcal{A}$ can then be ragar regarded as a kind of conjugate transpose of $\mathcal{A}$.

When $\mathcal{A}$ consists of one row and one column, having the single component a^{α}_{β}, I shall write simply

$$\mathcal{A} = a.$$

A 3.4a If $\mathcal{A}$ is any tensor matrix, a solution of the four equations $\mathcal{A}\cdot\mathcal{X}\cdot\mathcal{A}=\mathcal{A}$, $\mathcal{X}\cdot\mathcal{A}\cdot\mathcal{X}=\mathcal{X}$, $(\mathcal{A}\cdot\mathcal{X})^*=\mathcal{A}\cdot\mathcal{X}$, $(\mathcal{X}\cdot\mathcal{A})^*=\mathcal{X}\cdot\mathcal{A}$, will be called a generalized inverse of $\mathcal{A}$ and written $\mathcal{X}=\mathcal{A}^\dagger$, as in A 2.2. Now any tensor matrix $\mathcal{A}$ over $\mathfrak{C}$ can be regarded as the block matrix of the sets of components of the elements of $\mathcal{A}$ in some frame $\mathcal{F}$. Thus the existence and uniqueness of the generalized inverse for tensor matrices follows from its existence and uniqueness for ordinary matrices over $\mathfrak{C}_1$ and this follows from A 2.2.

A 3.4b Similarly, by A 2.21, we have $\mathcal{A}^\dagger=(\mathcal{A}^*\cdot\mathcal{A})^\dagger\cdot\mathcal{A}^*$.

A 3.4c Also, it follows from A 2.5 that the general solution of the tensor matrix equation of $\mathcal{A}\cdot\mathcal{X}=\mathcal{B}$, when one exists, is $\mathcal{X}=\mathcal{A}^\dagger\cdot\mathcal{B}+\mathcal{Y}-\mathcal{A}^\dagger\cdot\mathcal{A}\cdot\mathcal{Y}$ ($\mathcal{Y}$ arbitrary). Thus simultaneous linear equations in $\mathfrak{C}$, with several unknowns, may be treated in this way.

A 3.5 Theorem. Let $\mathscr{C}$ be normal and complete and let $\mathscr{S}$ be a normal subsystem of $\mathscr{C}$, let $a_{(1)\beta}^{\xi}, \ldots, a_{(\lambda)\beta}^{\theta}, b_{(1)\gamma}^{\xi}, \ldots, b_{(\lambda)\gamma}^{\theta} \in \mathscr{S}$; then the general solution in $\mathscr{S}$ of the set of simultaneous equations

$$a_{(1)} \cdot x = b_{(1)}, \ \ldots, \ a_{(\lambda)} \cdot x = b_{(\lambda)}$$

when one exists, is given by

$$x = c^{\dagger} \cdot (a_{(1)}^{*} \cdot b_{(1)} + \ldots + a_{(\lambda)}^{*} \cdot b_{(\lambda)}) + (I - c^{\dagger} \cdot c) \cdot y,$$

where $c = a_{(1)}^{*} \cdot a_{(1)} + \ldots + a_{(\lambda)}^{*} \cdot a_{(\lambda)}$

and where y_{γ}^{β} is an arbitrary element of $\mathscr{S}_{\beta * \gamma}$.

Proof. The given equations may be combined as the single tensor matrix equation $\mathscr{A} \cdot x = \mathscr{B}$ where

$$\mathscr{A} = \begin{pmatrix} a_{(1)} \\ \hline \vdots \\ \hline a_{(\lambda)} \end{pmatrix} \quad \text{and} \quad \mathscr{B} = \begin{pmatrix} b_{(1)} \\ \hline \vdots \\ \hline b_{(\lambda)} \end{pmatrix}.$$

The general solution of this equation, when one exists, is $x = \mathscr{A}^{\dagger} \cdot \mathscr{B} + y - \mathscr{A}^{\dagger} \cdot \mathscr{A} \cdot y$, by A 3.4c. But, by A 3.4b, $\mathscr{A}^{\dagger} = (\mathscr{A}^{*} \cdot \mathscr{A})^{\dagger} \cdot \mathscr{A}^{*}$. Now $\mathscr{A}^{*} \cdot \mathscr{A} = c$, so that the required solution is $x = c^{\dagger} \cdot \mathscr{A}^{*} \cdot \mathscr{B} + y - c^{\dagger} \cdot c \cdot y$ and the result follows.

A 3.5a Remark. Simultaneous linear equations like those of A 3.5 can also be treated in a regular system $\mathscr{S}$ which is not normal. Denoting by $p^{\vee}$ a solution in $\mathscr{S}$ of $p \cdot x \cdot p = p$ $(p \in \mathscr{S})$ it may be verified that if $\mathscr{A} = (b|c)$ then $\mathscr{A} \cdot \mathscr{A}^{\vee} \cdot \mathscr{A} = \mathscr{A}$ where $\mathscr{A}^{\vee} = \begin{pmatrix} x \\ \hline y \end{pmatrix}$, $x = ((I - c \cdot c^{\vee}) \cdot b)^{\vee} \cdot (I - c \cdot c^{\vee})$ and $y = c^{\vee} \cdot (I - b \cdot x)$. This can be extended to treat any tensor matrix whatever with elements in $\mathscr{S}$. This is the result that any matrix ring over

a regular ring is itself regular. See Brown and McCoy (3).) Thus a system of linear equations $a_{(i,1)} \cdot x_{(1)} + .. + a_{(i,\lambda)} \cdot x_{(\lambda)} = b_{(i)}$ $(i = 1, .., s)$ can be treated in a regular system in this way.

If $\mathcal{J}$ is a normal subsystem of $\mathcal{C}$, it can also be shown that the generalized inverse of any tensor matrix over $\mathcal{J}$ is explicitly expressible in terms of generalized inverses of elements of $\mathcal{J}$ and is thus a tensor matrix over $\mathcal{J}$. This is done in a particular case in the proof of A 3.5.

A 3.5b The simultaneous equations of A 3.5 may also be treated in the same way for elements of a group algebra since A 2.10h essentially shows that a group tensor system can be regarded as a normal subsystem of a complete system. This also enables us to treat linear equations in elements of $\mathcal{J}^{(\alpha)}$ since, according to A 1.9c, an element of $\mathcal{J}^{(\alpha)}$ can be associated uniquely with an element A of $\mathcal{J}$ for which $A \cdot Q^{(\lambda)}_{\nu(\alpha)} = 0$.

A 3.5c Lemma. If a^{α}_{ξ} , .. , c^{γ}_{ξ} , $p^{\xi}_{\eta} \in \mathcal{C}$ and $(a^* \cdot a + .. + c^* \cdot c) \cdot p = 0$, then $a \cdot p = 0$, .. , $c \cdot p = 0$.

Proof. Put $\mathcal{A} = \left(\frac{a}{\cdot\cdot}\right)$. Now $\mathcal{A}^* \cdot \mathcal{A} \cdot p = 0$ implies $\mathcal{A} \cdot p = 0$ by A 1.6m.

A 3.5d Lemma. Let a^{ξ}_{α} , b^{β}_{α} , .. , $d^{\delta}_{\alpha} \in \mathcal{C}$. Then a necessary and sufficient condition for there to exist x^{ξ}_{β} , .. , $z^{\xi}_{\delta} \in \mathcal{C}$ such that $a = x \cdot b + .. + z \cdot d$ is that for some $r \geqslant 0$, $\{a(b)^t .. (d)^v\} \cdot N^{(\lambda+1)} = 0$ for all $t, .., v \geqslant 0$ for which $t + .. + v = r$ and that for some such $t, .., v$, $\{(b)^t .. (d)^v\} \cdot N \neq 0$.

Proof. Put $\mathcal{B} = \left(\frac{b}{\cdot\cdot}\right)$. We require the condition for the existence of $\mathcal{X} = (x|..|z)$ such that $a = \mathcal{X} \cdot \mathcal{B}$, or equivalently, the existence of

$\mathcal{Y}(=\mathcal{X}\cdot\mathcal{B}^{\dagger *})$ such that $\mathfrak{a} = \mathcal{Y}\cdot(\mathcal{B}^{*}\cdot\mathcal{B})$ (since $\mathcal{B}^{\dagger *}\cdot\mathcal{B}^{*}\cdot\mathcal{B}=\mathcal{B}$, see A 2.2). Now $\mathcal{B}^{*}\cdot\mathcal{B}$ is the single element $b^{*}\cdot b + .. + d^{*}\cdot d$ so that A 2.6 applies. Thus the required condition is the existence of an $r \geqslant 0$ such that $(b^{*}\cdot b + .. + d^{*}\cdot d)^{\wedge}\cdot N^{(\wedge)} \neq 0$ but such that $\mathfrak{a}(b^{*}\cdot b + .. + d^{*}\cdot d)^{\wedge}\cdot N^{(\wedge+1)}\neq 0$.

Now, by A 1.8u and A 1.5p, $(b^{*}\cdot b + .. + d^{*}\cdot d)^{\wedge}\cdot N^{(\wedge)}$

$$= N^{(\wedge)}\cdot(b^{*}\cdot b + .. + d^{*}\cdot d)^{\wedge}\cdot N^{(\wedge)}$$

$$= \sum_{\xi+..+\upsilon=\wedge} \frac{\wedge!}{\xi!..\upsilon!}\, N^{(\wedge)}\cdot\left\{(b^{*}\cdot b)^{\xi} .. (d^{*}\cdot d)^{\upsilon}\right\}\cdot N^{(\wedge)}$$

$$= \sum_{\xi+..+\upsilon=\wedge} \frac{\wedge!}{\xi!..\upsilon!}\left\{(b)^{\xi} .. (d)^{\upsilon}\cdot N^{(\wedge)}\right\}^{*}\cdot\left\{(b)^{\xi} .. (d)^{\upsilon}\cdot N^{(\wedge)}\right\}$$

Also by A 1.5r,

$$\mathfrak{a}(b^{*}\cdot b + .. + d^{*}\cdot d)^{\wedge}\cdot N^{(\wedge+1)} = \left\{(b^{*}\cdot b + .. + d^{*}\cdot d)^{\wedge}\cdot N^{(\wedge)}\right\}\cdot(aI^{(\wedge)}\cdot N^{(\wedge+1)})$$

$$= \sum_{\xi+..+\upsilon=\wedge} \frac{\wedge!}{\xi!..\upsilon!}\left\{(b)^{\xi} .. (d)^{\upsilon}\cdot N^{(\wedge)}\right\}^{*}\cdot\left\{(b)^{\xi} .. (d)^{\upsilon}\cdot N^{(\wedge)}\right\}\cdot(aI^{(\wedge)}\cdot N^{(\wedge+1)}).$$

Hence the result follows from A 3.5c.

A 3.5e Lemma. If $x_{(1)\alpha}, .., x_{(\wedge)\alpha}$ span $\mathcal{C}_{\alpha}$ and a^{α}_{β}, b^{γ}_{β}, $y_{(1)\gamma}, .., y_{(\wedge)\gamma} \in \mathcal{C}$ and $x_{(i)}\cdot\mathfrak{a} = y_{(i)}\cdot b$ $(i = 1, .., r)$, then there exists $z^{\alpha}_{\alpha} \in \mathcal{C}$ such that $\mathfrak{a} = z\cdot b$.

Proof. We can choose $z^{\alpha}_{(1)}, .., z^{\alpha}_{(\wedge)}$ so that $\sum_{i=1}^{\wedge} z_{(i)}\, x_{(i)} = I_{(\alpha)}$ (see A 2.8a and A 3.1b). Put $z = \sum_{i=1}^{\wedge} z_{(i)}\, y_{(i)}$.

A 4 Tensors representing linear sets.

Linear subsets of a normal tensor system may be represented by single tensors in two different ways, namely by projectors or by Grassmann tensors.

Properties of these tensors and relations between the two types will be discussed in this chapter. The system $\mathcal{C} = \mathcal{C}(\mathfrak{G}, \mathfrak{R})$ will be supposed normal (i.e. $\mathfrak{R}$ is formally complex) and $\mathcal{J}$ will be a normal subsystem of $\mathcal{C}$.

A 4.1 Projectors.

A projector is an element of $\mathcal{C}$ which satisfies $e^* \bullet e = e$, that is $e \bullet e = e = e^*$.

The elements $I^{(\lambda)}$, $P^{(\lambda)}$, $N^{(\lambda)}$, $T^{\{\lambda\}}$, $G^{\{\lambda\}}$, $Q_m^{(\lambda)}$ of $\mathcal{R}$ are all projectors. If e is a projector then $e^\dagger = e$.

A 4.1a Lemma. If e and f are projectors, so are ef and I - e. If $a \in \mathcal{C}$, then $a^\dagger \bullet a$ and $a \bullet a^\dagger$ are projectors.

Proof. This follows from A 1.3k, A 1.6d and from A 2.2.

A 4.1b Definition. An element x_α^γ of $\mathcal{C}$ satisfying the relation $x \bullet a = 0$, where $a_\beta^\alpha \in \mathcal{J}$, is called a left annihilator of a. Clearly, the left annihilators of a in $\mathcal{C}_{\beta^* \gamma}$ of in $\mathcal{J}_{\beta^* \gamma}$ form a linear set over $\mathcal{C}_1$ or $\mathcal{J}_1$ respectively. Right annihilators are similarly defined.

A 4.1c Lemma. Let $a_\beta^\alpha \in \mathcal{J}$; then the projector $a \bullet a^\dagger$ [the projector $a^\dagger \bullet a$] has the same set of left [right] annihilators in $\mathcal{C}$ as a has.

Proof. This follows from the relation $a \bullet a^\dagger \bullet a = a$.

A 4.1d Lemma. Let e_κ^α, f_α^α be projectors in $\mathcal{J}$; then the space of left [right] annihilators, in $\mathcal{C}_\alpha$ [in $\mathcal{C}_{\kappa^*}$] or in $\mathcal{J}_{\kappa\kappa^*}$, of e, is included in that of f, if and only if $e \bullet f = f$ [$f \bullet e = f$].

CHAPTER 6

Proof. The tensor I - e is a left annihilator, in $\mathcal{S}_{\alpha\alpha^*}$, of e. Thus, if I - e is also a left annihilator of f, we have e•f = f. Also if f - e•f ≠ 0, there exists an element x_α of $\mathcal{C}_\alpha$ such that x•(f - e•f) ≠ 0, since $\mathcal{C}$ is complete, so that x•(I - e), though a left annihilator in $\mathcal{C}_\alpha$ of e, is not a left annihilator of f.

Conversely, if e•f = f, then clearly, any left annihilator of e, whether in $\mathcal{C}_\alpha$ or in $\mathcal{S}_{\alpha\alpha^*}$, is also a left annihilator of f.

The proof for right annihilators is similar.

A 4.1e Lemma. Let e^α_α, f^α_α be projectors in $\mathcal{S}$; then e•f = f implies f•e = f.

Proof. If e•f = f, then f•e = f^*•e^* = (e•f)* = f^* = f.

A 4.2 Theorem. Any projector e^α_α of $\mathcal{S}$ is uniquely determined by its space of left [right] annihilators in $\mathcal{C}_\alpha$ [$\mathcal{C}_{\alpha^*}$] or in $\mathcal{S}_{\alpha\alpha^*}$.

Proof. Let e, f be two projectors in $\mathcal{S}$ with the same space of left annihilators in $\mathcal{C}_\alpha$ or in $\mathcal{S}_{\alpha\alpha^*}$. By A 4.1d, we have e•f = f and f•e = e. The result follows from A 4.1e. A similar argument holds for right annihilators.

A 4.2a The dimensions of the left [right] annihilator space in $\mathcal{C}_\alpha$ [$\mathcal{C}_{\alpha^*}$] and in $\mathcal{S}_\alpha$ [$\mathcal{S}_{\alpha^*}$] of the projector e^α_α of $\mathcal{S}$, are respectively $\nu(\alpha) - \langle e \rangle$ and $\langle u_{(\alpha)} \rangle - \langle u_{(\alpha)} \bullet e \rangle$, by A 2.9c.

A 4.2b Definition. Let $e^\alpha_{(1)\alpha}$, ..., $e^\alpha_{(\wedge)\alpha}$ be projectors in $\mathcal{S}$. I define the intersection of $e_{(1)}$, .., $e_{(\wedge)}$ to be the projector

$$[e_{(1)}, \ldots, e_{(\lambda)}] = (e_{(1)} + \ldots + e_{(\lambda)})^{\dagger} \cdot (e_{(1)} + \ldots + e_{(\lambda)})$$
$$= (e_{(1)} + \ldots + e_{(\lambda)}) \cdot (e_{(1)} + \ldots + e_{(\lambda)})^{\dagger}$$

(by A 2.2n) and the <u>join</u> of $e_{(1)}, \ldots, e_{(\lambda)}$ to be the projector

$$(e_{(1)}, \ldots, e_{(\lambda)}) = I - [I - e_{(1)}, \ldots, I - e_{(\lambda)}].$$

The join and intersection of $e_{(1)}, \ldots, e_{(\lambda)}$ are obviously symmetrical in the $e_{(i)}$ and with respect to left and right contracted products.

A 4.2c <u>Lemma</u>. Let $e^{\alpha}_{(1)\alpha}, \ldots, e^{\alpha}_{(\lambda)\alpha}$ be projectors in $\mathcal{S}$; then if $x \in \mathcal{C}$ we have, $e_{(1)} \cdot x = \ldots = e_{(\lambda)} \cdot x = 0$ if and only if $[e_{(1)}, \ldots, e_{(\lambda)}] \cdot x = 0$. Also $e_{(1)} \cdot x = \ldots = e_{(\lambda)} \cdot x = x$ if and only if $(e_{(1)}, \ldots, e_{(\lambda)}) \cdot x = x$. The corresponding results also hold for $x \cdot e_{(1)}$, .. etc.

<u>Proof</u>. In theorem A 3.5, put $a_{(1)} = e_{(1)}, \ldots, a_{(\lambda)} = e_{(\lambda)}$ and $b_{(1)} = 0, \ldots, b_{(\lambda)} = 0$. We then obtain the result that those elements x of which satisfy all of the relations $e_{(1)} \cdot x = 0, \ldots, e_{(\lambda)} \cdot x = 0$ are the elements of the form $(I - c^{\dagger} \cdot c) \cdot y$ where $c = (e^{*}_{(1)} \cdot e_{(1)} + \ldots + e^{*}_{(\lambda)} \cdot e_{(\lambda)})$ $= (e_{(1)} + \ldots + e_{(\lambda)})$. Now the elements of the form $(I - c^{\dagger} \cdot c) \cdot y$ are those elements x which satisfy $c^{\dagger} \cdot c \cdot x = 0$ as in A 2.5. Thus the result follows by A 4.2b.

A 4.2d <u>Lemma</u>. $[e_{(1)}, \ldots, e_{(\lambda)}] \cdot e_{(i)} = e_{(i)} = e_{(i)} \cdot [e_{(1)}, \ldots, e_{(\lambda)}]$ and $(e_{(1)}, \ldots, e_{(\lambda)}) \cdot e_{(i)} = (e_{(1)}, \ldots, e_{(\lambda)}) = e_{(i)} \cdot (e_{(1)}, \ldots, e_{(\lambda)})$, $i = 1, \ldots, r$.

<u>Proof</u>. Putting $x = I - [e_{(1)}, \ldots, e_{(\lambda)}]$ in the first part of A 4.2c we obtain $e_{(i)} = e_{(i)} \cdot [e_{(1)}, \ldots, e_{(\lambda)}]$, whence by A 4.1e or by symmetry,

$e_{(i)} = [e_{(1)}, \ldots, e_{(\lambda)}] \cdot e_{(i)}$. The rest of the result follows from A 4.2b or by putting $x = (e_{(1)}, \ldots, e_{(\lambda)})$ in the second part of A 4.2c.

A 4.3 Theorem. Let $e_{(1)\alpha}^{\alpha}, \ldots, e_{(\lambda)\alpha}^{\alpha}$ be projectors of $\mathcal{S}$ and let $\gamma \in \mathfrak{G}$; then the spaces of left [right] annihilators, in $\mathcal{C}_{\alpha\gamma}$ [$\mathcal{C}_{\alpha^*\gamma}$], or in $\mathcal{S}_{\alpha\alpha^*}$ of $[e_{(1)}, \ldots, e_{(\lambda)}]$ and of $(e_{(1)}, \ldots, e_{(\lambda)})$ are, respectively, the intersection and join of those of $e_{(1)}, \ldots, e_{(\lambda)}$.

Proof. That the intersection of the spaces of right annihilators in $\mathcal{C}_{\alpha^*\gamma}$ of $e_{(1)}, \ldots, e_{(\lambda)}$ is the space of right annihilators in $\mathcal{C}_{\alpha^*\gamma}$ of $[e_{(1)}, \ldots, e_{(\lambda)}]$, is simply the first part of A 4.2c.

Now let $e_{(i)} \cdot y_{(i)} = 0$, $i = 1, \ldots, r$ where $y_{(i)\gamma}^{\alpha} \in \mathcal{C}_{\alpha^*\gamma}$; then $(e_{(1)}, \ldots, e_{(\lambda)}) \cdot y_{(i)} = (e_{(1)}, \ldots, e_{(\lambda)}) \cdot e_{(i)} \cdot y_{(i)} = 0$, by A 4.2d. Thus $(e_{(1)}, \ldots, e_{(\lambda)}) \cdot \{\lambda_1 y_{(1)} + \ldots + \lambda_\lambda y_{(\lambda)}\} = 0$, where $\lambda_1, \ldots, \lambda_\lambda \in \mathcal{C}_1$. Conversely, since $(e_{(1)}, \ldots, e_{(\lambda)}) = I - \sum_{i=1}^{\lambda} (I - e_{(i)}) \cdot (rI - e_{(1)} - \ldots - e_{(\lambda)})^{\dagger}$, if $(e_{(1)}, \ldots, e_{(\lambda)}) \cdot x = 0$, then $x = \sum_{i=1}^{\lambda} x_{(i)}$ where $x_{(i)} = (I - e_{(i)}) \cdot (rI - e_{(1)} - \ldots - e_{(\lambda)})^{\dagger} \cdot x$ so that $e_{(i)} \cdot x_{(i)} = 0$. Thus, the join of the spaces of right annihilators in $\mathcal{C}_{\alpha^*\gamma}$ of $e_{(1)}, \ldots, e_{(\lambda)}$ is the space of right annihilators in $\mathcal{C}_{\alpha^*\gamma}$ of $(e_{(1)}, \ldots, e_{(\lambda)})$. The result for left annihilator spaces follows by symmetry.

A 4.3a By A 4.1d and A 4.3, the relation $e \cdot f = e$ is equivelent to $(e, f) = e$ and to $[e, f] = f$.

A 4.3b Remark. If $\mathcal{S}$ is a regular, but not a normal system, then a projector e of $\mathcal{S}$ could be defined simply as an element of $\mathcal{S}$ such that

$e \cdot e = e$. The results of A 4.1c and A 4.1d would remain, but not A 4.1e or the uniqueness theorem A 4.2. The left intersection of two projectors e and f could be defined as $g = e + (I - e) \cdot f \cdot ((I - e) \cdot f)^{\nu} \cdot (I - e)$ where, as in A 3.5a, p^{ν} is a solution of $p \cdot x \cdot p = p$; then g is a projector such that $x \cdot g = 0$ if and only if $x \cdot e = 0 = x \cdot f$. However, such a projector is not unique and, in fact the definition of g given, is not symmetrical in e and f. There is the added disadvantage that a left intersection of projectors is not necessarily also a right intersection. Also, in a practical case, finding the left intersection of r projectors would involve solving an equation such as $p \cdot x \cdot p = p$, $r - 1$ times. This need only be done once (and A 2.4 may be used) in a normal system. The simplification of the theory that is obtained by taking $\mathcal{S}$ to be normal, is, therefore, considerable.

The formula given above, for the left intersection of two projectors (when $\mathcal{S}$ is regular but not normal), is adapted from a result due to Rutherford (20) (pp 81-88) on the solution of simultaneous linear substitutional equations and it is closely related to the formulae of A 3.5a. However, Rutherford never explicitly mentions the use of a solution of $p \cdot x \cdot p = p$. He calls a right intersection of projectors a "master idempotent."

A 4.3c Remark. Miscellaneous formulae involving two projectors e, f of a normal system $\mathcal{S}$ may be given as follows :

$$[e,f] = e + (I-e)\cdot f\cdot((I-e)\cdot f)^{\dagger}\cdot(I-e) = e + (I-e)\cdot((I-e)\cdot f)^{\dagger}$$
$$= (e+f-e\cdot f)\cdot(I - (e\cdot f-f\cdot e)^{\dagger}\cdot(e\cdot f-f\cdot e)) = e + (I-e)\cdot(e-f)^{\dagger}\cdot(e-f)$$
$$= e - (I-e)\cdot(e-f)^{\dagger}\cdot(I-e) = (e+f)\cdot(e+f)^{\dagger},$$

$$(e,f) = e - e\cdot(e\cdot(I-f))^{\dagger} = e\cdot f\cdot(I - (e\cdot f-f\cdot e)^{\dagger}\cdot(e\cdot f-f\cdot e))$$
$$= I - (I-e\cdot f)^{\dagger}\cdot(I-e\cdot f) = e\cdot(I - (e-f)^{\dagger}\cdot(e-f)) = e - e\cdot(e-f)^{\dagger}\cdot e.$$

Other relations are obtained by interchanging e and f or by reversing the order of the products.

A 4.3d Since the intersection or join of any (not necessarily finite) number of linear spaces is again a linear space, the definitions A 4.2b of the intersection and join of a finite number of projectors can be extended to treat an infinite number of projectors. If $\mathscr{E}$ is any set of projectors e_{κ}^{α} of $\mathcal{S}_{\alpha\alpha^*}$, the intersection and join of the projectors e may be denoted, respectively, by

$$[\, e\,;\ e\in\mathscr{E}\,] \quad \text{and} \quad (e\,;\ e\in\mathscr{E}\,).$$

These projectors are uniquely defined (see A 4.2) by the fact that their right (or left) annihilator spaces in $\mathcal{S}_{\alpha\alpha^*}$ or in $\mathcal{C}_{\alpha\gamma}$ (or $\mathcal{C}_{\alpha^*\gamma}$) are, respectively, the join and intersection of those of the projectors e of $\mathscr{E}$. The phrase "$e\in\mathscr{E}$" following the semicolon in the notation may be replaced by any other phrase or set of phrases characterizing the set of projectors concerned.

A 4.4 Grassmann tensors

An α-Grassmann tensor is defined as an element $p^{\alpha_1 \cdots \alpha_\lambda}$ of $\mathcal{C}$ which is of the form

$$p = N^{(\lambda)}_{\bullet}\, a_{(1)} \,..\, a_{(\lambda)}$$

where $a^{\alpha}_{(1)}, \ldots, a^{\alpha}_{(\lambda)} \in \mathcal{C}_{\alpha^*}$. If $q_{\alpha_1^* \cdots \alpha_\lambda^*} = p^{\alpha_1 \cdots \alpha_\lambda}$, then q is called a dual α^*-Grassmann tensor. (The difference is purely a notational one.)

A 4.4a Using A 1.5m we can write

$$p^{\alpha_1 \cdots \alpha_\lambda} = \frac{1}{\lambda!} \begin{vmatrix} a^{\alpha_1}_{(1)} & \cdots & a^{\alpha_\lambda}_{(1)} \\ \vdots & & \vdots \\ a^{\alpha_1}_{(\lambda)} & \cdots & a^{\alpha_\lambda}_{(\lambda)} \end{vmatrix} .$$

It follows that the components of p in any frame $\mathcal{F}$ are given by a similar determinant in the components in the $a_{(i)}$. Hence the components of p in $\mathcal{F}$ form a set of Grassmann coordinates (see Hodge and Pedoe (9) pp.288-290). Grassmann coordinates are sometimes written with lower indices and dual Grassmann coordinates with upper indices.

A 4.4b Two elements a and b of $\mathcal{C}$ will be called proportional if there is a non-zero scalar λ such that $a = \lambda b$.

A 4.4c For each $\alpha \in \mathfrak{G}$ there is a one to one correspondence between the linear sets in $\mathcal{C}_{\alpha^*}$ (over $\mathcal{C}_1$) and the equivalence classes of proportional α-Grassmann tensors in $\mathcal{C}$. This follows from the corresponding property for Grassmann coordinates (see (9) pp.289-290). If the linear set is spanned by the vectors $a^{\alpha}_{(1)}, \ldots, a^{\alpha}_{(\lambda)}$, a corresponding Grassmann tensor is given by $p = N^{(\lambda)}_{\bullet}\, a_{(1)} \,..\, a_{(\lambda)}$.

A 4.4d Linear subsets of $\mathcal{C}_{\alpha^*}$ may also be represented by dual α-Grassmann tensors. Thus, if q is a dual α-Grassmann tensor and $q_{\alpha_1 \cdots \alpha_\lambda} = p^{\alpha_1^* \cdots \alpha_\lambda^*}$, then the linear subset of $\mathcal{C}_{\alpha^*}$ represented by the dual α-Grassmann tensor q is the dual space to the linear subset of $\mathcal{C}_\alpha$ represented by the α^*-Grassmann tensor p. (The dual space of a linear subset $\mathcal{L}$ of $\mathcal{C}_\alpha$ is the set of elements x^α of $\mathcal{C}_{\alpha^*}$ satisfying $a \cdot x = 0$ for all $a_\alpha \in \mathcal{L}$.) Relations between α-Grassmann tensors and the corresponding dual α-Grassmann tensors will be given in A 4.5d.

A 4.4e It is usual to associate Grassmann coordinates (or dual Grassmann coordinates) with Projective linear spaces rather than linear vector spaces. Thus a Grassmann tensor $p^{\alpha_1 \cdots \alpha_\lambda}$ corresponding to a linear subset $\mathcal{L}$ of $\mathcal{C}_{\alpha^*}$ may be associated with the projective linear (r-1)-space whose points are the equivalence classes of proportional elements of $\mathcal{L}$. It will be convenient to denote the projective space associated with the tensor p by the same symbol p and it is hoped that this will not lead to confusion. Similarly, if $q_{\alpha_1 \cdots \alpha_s}$ is a dual Grassmann tensor it is associated with a projective linear $(\nu(\alpha)-1-s)$-space and this space will sometimes be denoted by q.

A 4.4f Lemma. Let p^{α_λ}, p'^{α_μ} be α-Grassmann tensors. Then the join of the spaces p and p′ is represented by the α-Grassmann tensor $N^{(\lambda+\mu)} \cdot pp'$ unless the spaces p and p′ meet, in which case $N^{(\lambda+\mu)} \cdot pp' = 0$.
Similarly, if q_{α^λ}, q'_{α^μ} are dual α-Grassmann tensors the intersection of

the spaces q, q′ is given by $qq' \cdot N^{(\lambda+\delta)}$ except that $qq' \cdot N^{(\lambda+\delta)} = 0$ if there is a prime containing both q and q′.

Proof. Let $p = N^{(\lambda)} \cdot a_{(1)} \ldots a_{(\lambda)}$ and $p' = N^{(\delta)} \cdot a_{(\lambda+1)} \ldots a_{(\lambda+\delta)}$. Then $N^{(\lambda+\delta)} \cdot pp' = N^{(\lambda+\delta)} \cdot \{N^{(\lambda)} \cdot a_{(1)} \ldots a_{(\lambda)}\}\{N^{(\delta)} \cdot a_{(\lambda+1)} \ldots a_{(\lambda+\delta)}\}$ $= N^{(\lambda+\delta)} \cdot a_{(1)} \ldots a_{(\lambda+\delta)}$ by A 1.5r. Thus the result for Grassmann tensors follows from A 4.4c and A 2.7. The result for dual Grassmann tensors follows from this by duality.

A 4.4g **Lemma.** Let p^{α_λ} be a Grassmann tensor and q_{α_λ} a dual Grassmann tensor. Then a necessary and sufficient condition for p to meet q in a space of at least t-r dimensions is $qI^{(t-\lambda)} \cdot I^{(t-\lambda)} p = 0$.

Proof. This result is given in Hodge and Pedoe (9) p.304.

A 4.4h **Lemma.** Let p^{α_λ} be a Grassmann tensor and q_{α_δ} a dual Grassmann tensor. Then if p does not meet q, a dual Grassmann tensor representing their join is $q \cdot I^{(\lambda-\delta)} p$. Dually, if there is no prime containing both p and q, their intersection is given by $qI^{(\delta-\lambda)} \cdot p$ $(= \pm I^{(\delta-\lambda)} q \cdot p)$.

Proof. This result (and also A 4.4f) is given in (9) p.307.

A 4.4i **Lemma.** If p^{α_λ} and p'^{α_δ} be Grassmann tensors, then $I^{(\lambda-1)} N^{(\delta+1)} \cdot pp'$ $= 0$ if and only if the space p is contained in p′. There is also a corresponding dual result.

Proof. Let $p = N^{(\lambda)} \cdot a_{(1)} \ldots a_{(\lambda)}$ and $p' = N^{(\delta)} \cdot b_{(1)} \ldots b_{(\delta)}$. Suppose the space p is contained in p′. Then each of $a_{(1)}, \ldots, a_{(\lambda)}$ is linearly

CHAPTER 6

dependent on $b_{(1)}, \ldots, b_{(\lambda)}$ so we have $0 = N^{(\lambda+1)}_{\bullet} a_{(i)}\, b_{(1)} \ldots b_{(\lambda)}$ $= N^{(\lambda+1)}_{\bullet} a_{(i)} p'$, by A 2.7 and A 1.5r. Hence $I^{(\lambda-1)} N^{(\lambda+1)} \bullet pp'$ $= I^{(\lambda-1)} N^{(\lambda+1)}_{\bullet} \{N^{(\lambda)}_{\bullet} a_{(1)} \ldots a_{(\lambda)}\} p' = \sum \pm a_{(i_1)} \ldots a_{(i_{\lambda-1})} (N^{(\lambda+1)}_{\bullet} a_{(i_\lambda)} p') = 0$. Conversely, if $I^{(\lambda-1)} N^{(\lambda+1)} \bullet pp' = 0$ and $q_{\alpha^{\lambda-1}}$ is a dual Grassmann tensor of an arbitrary projective $(n-\lambda)$-space, then $0 = qN^{(\lambda+1)}_{\bullet} pp'$ $= N^{(\lambda+1)} \bullet (I^{(1)} q \bullet p)p'$, so that by A 4.4h and A 4.4f, the intersection of q and p is a point of p', when the space p' contains p.

A 4.4j Lemma. If p^{α^λ} is a Grassmann tensor and q_{α^λ} is a dual Grassmann tensor then $I^{(\lambda-1)} N^{(\lambda+1)} \bullet (p)^2 = 0$ and $(q)^2 \bullet I^{(\lambda-1)} N^{(\lambda+1)} = 0$.

Proof. This follows from A 4.4i since any space is contained in itself. These relations are, in fact, equivalent to the quadratic p-relations for Grassmann coordinates (see (9) pp.309-314). A more usual form for these relations is $\sum_{i=1}^{\lambda+1} (-1)^i p^{\alpha_1 \ldots \alpha_{\lambda-1} \alpha_{\lambda+i-1}} p^{\alpha_\lambda \ldots \alpha_{\lambda+i-2} \alpha_{\lambda+i} \ldots \alpha_{i\lambda}} = 0$. The equivalence of this to the relations given above follows from A 1.5s and the skew-symmetry of p.

A 4.4k Lemma. If p^{α^λ} is a Grassmann tensor and q_{α^λ} is a dual Grassmann tensor, then $I^{(\lambda-1)} qI^{(1)} \bullet (p)^2 = \frac{(-1)^{\lambda-1}(q \bullet p)}{\lambda} p$ and $(q)^2 \bullet I^{(\lambda-1)} pI^{(1)}$ $= \frac{(-1)^{\lambda-1}(q \bullet p)}{\lambda} q$.

Proof. From A 4.4j we get $I^{(\lambda)} q \bullet I^{(\lambda-1)} N^{(\lambda+1)} \bullet (p)^2 = 0$ whence by A 1.5s and the skew-symmetry of $N^{(\lambda)}$ we get

$\frac{1}{\lambda+1} I^{(\lambda)} q \bullet I^{(\lambda)} N^{(\lambda)} \bullet (p)^{s} = (-1)^{\lambda-1} \frac{\lambda}{\lambda+1} I^{(\lambda-1)} q I^{(1)} \bullet I^{(\lambda-1)} N^{(\lambda)} I^{(1)} \bullet I^{(\lambda)} N^{(\lambda)} \bullet (p)^{s}$ and the result follows since $q \bullet N^{(\lambda)} = q$ and $N^{(\lambda)} \bullet p = p$.

A 4.4l Lemma. If p and q are as in A 4.4k, $q \bullet p \neq 0$ and $e_{\alpha}^{\alpha} = \frac{(-1)^{\lambda-1}\lambda}{(q \bullet p)} I^{(\lambda)} q \bullet p I^{(1)}$ then $e \bullet e = e$.

Proof. This follows at once from A 4.4k since $I^{(\lambda)} q \bullet p I^{(1)} = q I^{(1)} \bullet I^{(1)} p$ by skew-symmetry of p and q.

A 4.4m Lemma. If $e_{\alpha}^{\alpha} \in \mathcal{C}$ is an idempotent and $\langle e \rangle = m$, then

$(P^{(\lambda)} \bullet (e)^{\lambda})_{[\alpha^{\lambda-t}]} = \frac{(m+\lambda-1)(m+\lambda-2) \cdots (m+t)}{\lambda(\lambda-1) \cdots (t+1)} P^{(t)} \bullet (e)^{t}$ and

$(N^{(\lambda)} \bullet (e)^{\lambda})_{[\alpha^{\lambda-t}]} = \frac{(m-\lambda+1)(m-\lambda+2) \cdots (m-t)}{\lambda(\lambda-1) \cdots (t+1)} N^{(t)} \bullet (e)^{t}$.

Proof. This result is exactly analogous to A 1.5u but with the idempotent e replacing the unit $I_{(\alpha)}$. The proof is similar to that of A 1.5u since $N^{(m+1)} \bullet (e)^{m+1} = 0$ (see A 2.3c).

A 4.4n Lemma. Let e_{α}^{α} be idempotent with $\langle e \rangle = m$. Let q be a dual α-Grassmann tensor corresponding to the right annihilator space of e and let p be an α-Grassmann tensor corresponding to the dual space of the left annihilator space of e. Then

$$N^{(m)} \bullet (e)^{m} = \frac{1}{(q \bullet p)} pq$$

and e, p and q are related as in A 4.4l.

Proof. Since $e \bullet e = e$, the right annihilator space $\mathcal{L}$ of e does not meet the dual space (see A 4.4d) of the left annihilator space of e (nexcept at o^{α}). Thus, by A 4.4g, $q \bullet p \neq 0$. Also, by A 4.2a, the dimension of $\mathcal{L}$ is m. The right annihilators of e are all right annihilators of $e' = \frac{(-1)^{m-1} m}{(q \bullet p)} I^{(1)} q \bullet p I^{(1)}$ since by A 4.4g (with $m = t = r$, $s = 1$), $q \bullet I^{(m-1)} x = 0$ whenever $x^{\alpha} \in \mathcal{L}$. Hence the right

annihilator spaces of e and e′ are the same. By a corresponding dual argument (see A 4.4) the left annihilator spaces of e and e′ are also the same. Thus $e\bullet(I-e') = 0$ and $(I-e)\bullet e' = 0$ whence $e = e'$. By A 4.4m $\langle N^{(m)}\bullet(e)^m\rangle = 1$. Also, $N^{(m)}\bullet(e)^m$ is an idempotent by A 1.8u . Hence, by A 2.3c, the rank of $N^{(m)}\bullet(e)^m$ is unity, so that its matrix of components in any frame is the outer product of two non-zero vectors, whence $N^{(m)}\bullet(e)^m = uv$ where u^{α^m}, $v_{\alpha^m} \in \mathcal{C}$. Let $p = N^{(m)}\bullet a_{(i_1)}..a_{(i_m)}$ and $q = b_{(i_1)}..b_{(i_m)}\bullet N^{(m)}$, then e is of the form $e = \sum \lambda_{ij}\, a_{(i)}\, b_{(j)}$. Now if $N^{(m)}\bullet(e)^m = uv$, then u and v must clearly be both α-skew-symmetric. Hence u and v are of the form $u = \sum\mu N^{(m)}\bullet a_{(i_1)} .. a_{(i_m)}$, $v = \sum\mu\, b_{(j_1)} .. b_{(j_m)}\bullet N^{(m)}$. It follows that u is proportional to p and v to q (by the skew-symmetry of $N^{(m)}$) and the result follows since $\langle N^{(m)}\bullet(e)^m\rangle = 1$.

The first part of A 4.4n in fact does not depend on e being an idempotent provided that m is defined as its rank and $q\bullet p \neq 0$.

A 4.4o The result of A 4.4n enables us to obtain the dual Grassmann tensors corresponding to a linear subset $\mathcal{L}$ of $\mathcal{C}_{\alpha^*}$ from the projector e which has $\mathcal{L}$ as its right annihilator space. The required dual Grassmann tensor can be obtained in the form $a\bullet N^{(\lambda)}\bullet(e)^\lambda$, where $a_{\alpha^\lambda} \in \mathcal{C}$ is chosen so that this does not vanish.

In a similar way it can be shown that the Grassmann tensors corresponding to $\mathcal{L}$ can be obtained in the form $N^{(m-\lambda)}\bullet(I-e)^{m-\lambda}\bullet b$, where $b^{\alpha^{m-\lambda}} \in \mathcal{C}$ is chosen so that this does not vanish.

A 4.4p It is also possible to obtain the projector e^{α}_{α} from the corresponding dual Grassmann tensor $q_{\alpha_1 \cdots \alpha_\lambda}$. For putting $q_{(1)}{}^{\alpha_1^* \cdots \alpha_{\lambda-1}^*}{}_{\alpha_\lambda} = q_{\alpha_1 \cdots \alpha_\lambda}$ we have $q_{(1)} \bullet q^*_{(1)} \bullet q_{(1)} = \frac{1}{\lambda}\langle q^*_{(1)} \bullet q_{(1)}\rangle\, q_{(1)}$ by A 4.4k and the skew-symmetry of q whence $q^{\dagger}_{(1)} = \frac{\lambda}{\langle q^*_{(1)} \bullet q_{(1)}\rangle}\, q^*_{(1)}$ by A 2.2o. The right annihilator space of e is the same as that of $q_{(1)}$ (by A 4.4g with $t=r, s=1$) so that, by A 4.1c we have $e = q^{\dagger}_{(1)} \bullet q_{(1)}$, that is, $e = \frac{\lambda}{\langle q^*_{(1)} \bullet q_{(1)}\rangle}\, q^*_{(1)} \bullet q_{(1)}$

Similarly, if p is a Grassmann tensor corresponding to e and $p_{(1)}{}^{\alpha_1}{}_{\alpha_1^* \cdots \alpha_\lambda^*} = p^{\alpha_1 \cdots \alpha_\lambda}$, it can be shown that $e = I - \frac{\lambda}{\langle p_{(1)} \bullet p^*_{(1)}\rangle}\, p_{(1)} \bullet p^*_{(1)}$.

A 4.5 Alternating tensors

An α-Grassmann tensor $\epsilon^{(\alpha)} = \epsilon^{(\alpha)\alpha_1 \cdots \alpha_\lambda}$ corresponding to the space $\mathcal{C}_{\alpha^*}$ itself, will be called an α-alternating tensor. I shall also write $\epsilon_{(\alpha)}$ for $\epsilon_{(\alpha)\alpha_1 \cdots \alpha_\lambda} = \epsilon^{(\alpha^*)\alpha_1^* \cdots \alpha_\lambda^*}$.

A 4.5a If α is a generator of $\mathfrak{S}$, the tensors $\epsilon^{(\alpha)}$ and $\epsilon_{(\alpha)}$ may be chosen so that their components in a frame $\mathcal{F}$ are given by

$$\epsilon_{(\alpha) i_1 \cdots i_\lambda} = \epsilon^{(\alpha) i_1 \cdots i_\lambda} = \begin{cases} +1 \text{ if } (i_1, \ldots, i_\lambda) \text{ is an even permutation of } (1, \ldots, r) \\ -1 \quad \text{- - - - - - - odd - - - - - - -} \\ 0 \text{ otherwise} \end{cases}$$

(see Hodge and Pedoe (9) p.73). In this case $\epsilon_{(\alpha)} = (\epsilon^{(\alpha)})^*$ and I shall often write $\epsilon^{(\alpha)}$ and $\epsilon_{(\alpha)}$, respectively, as ϵ and ϵ^*.

A 4.5b From A 4.4n with $e = I_{(\alpha)}$ we see that $N^{(\lambda)}_{(\alpha)}$ is proportional to $\epsilon^{(\alpha)}\epsilon_{(\alpha)}$. The tensors $\epsilon^{(\alpha)}$ and $\epsilon_{(\alpha)}$ will always be chosen so that $\epsilon^{(\alpha)}\epsilon_{(\alpha)} = \nu(\alpha)!\, N^{(\nu(\alpha))}_{(\alpha)}$ (this is in agreement with A 4.5a). Hence, from

A 1.5u and the skew-symmetry of $\epsilon^{(\alpha)}$ we have

$$I^{(\lambda)}\epsilon_{(\alpha)}\cdot\epsilon^{(\alpha)}I^{(\lambda)} = (-1)^{\lambda(\nu(\alpha)-\lambda)}\, r!(\nu(\alpha)-\lambda)!\, N^{(\lambda)}. \qquad \text{A } 4.5c$$

A 4.5d If $p^{\alpha\hat{}}$ and $q_{\alpha^{\nu(\alpha)-\lambda}}$ are respectively a Grassmann tensor and a dual Grassmann tensor corresponding to the same space, then by A 4.4h, p is proportional to $qI^{(\lambda)}\cdot\epsilon^{(\alpha)}$. For convenience I shall always choose such associated Grassmann and dual Grassmann tensors to be related by

$$p = \frac{1}{\lambda!}\, qI^{(\lambda)}\cdot\epsilon^{(\alpha)}$$

since then, dually

$$q = \frac{1}{(m-\lambda)!}\,\epsilon_{(\alpha)}\cdot I^{(m-\lambda)}p$$

by A 4.5c.

A 4.5e Definition. Define the permutation $D^{(\ell:a;k)}$ by

$$D^{(\ell:a;k)} = J_{[1,\cdots,\ell,\ \ell+a+1,\cdots,2\ell+a,\ \cdots\cdots,\ (k-1)(\ell+a)+1,\cdots,k\ell+(k-1)a,\ \ell+1,\cdots,\ell+a,\ \cdots\cdots,\ k\ell+(k-1)a+1,\cdots,k\ell+ka]}.$$

Then if $a_{(i)} = a_{(i)\alpha_1\cdots\alpha_\ell}$, $b_{(i)} = b_{(i)\alpha_1\cdots\alpha_a}$, we have

$$a_{(1)}\cdots a_{(k)}\, b_{(1)}\cdots b_{(k)}\cdot D^{(\ell:a;k)} = a_{(1)}b_{(1)}a_{(2)}b_{(2)}\cdots a_{(k)}b_{(k)}$$

More particularly if $a = a_{(1)}$ and $b = b_{(1)}$, then

$$\begin{aligned} (a)^k\, I^{(ka)}\cdot D^{(\ell:a;k)} &= (aI^{(a)})^k, \\ I^{(k\ell)}(b)^k\cdot D^{(\ell:a;k)} &= (I^{(\ell)}b)^k. \end{aligned} \qquad \text{A } 4.5f$$

Hence, if p and q are as in A 4.5d and $\nu(\alpha) = n$, we have

$$\begin{aligned} (p)^a &= \frac{1}{(\lambda!)^a}(q)^a I^{(na)}\cdot D^{(n-\lambda:\lambda;a)}\cdot(\epsilon^{(\alpha)})^a, \\ (q)^a &= \frac{1}{\{(n-\lambda)!\}^a}(\epsilon_{(\alpha)})^a\, D^{(n-\lambda:\lambda;a)*}\cdot I^{(na-\lambda a)}(p)^a \end{aligned} \qquad \text{A } 4.5g$$

Also, since $I^{(1)}\epsilon_{(\alpha)}\cdot\epsilon^{(\alpha)}I^{(1)} = (-1)^{n-1}(n-1)!\, I^{(1)}$ by A 4.5c, we have

$$\frac{(-1)^{k(n-1)}}{\{(n-1)!\}^k}\, I^{(k)}\left\{(\epsilon_{(\alpha)})^k\cdot D^{(n-1:1;k)*}\right\}\cdot\left\{D^{(1:n-1;k)}\cdot(\epsilon^{(\alpha)})^k\right\}I^{(k)} = I^{(k)}. \qquad \text{A } 4.5h$$

Section B : Some applications of tensor techniques to algebraic geometry.

This section will be concerned throughout with the complete tensor system $\mathscr{C} = \mathscr{C}(\mathfrak{S}, \mathfrak{R})$, where $\mathfrak{S}$ is semigenerated by a single index type α, with $\nu(\alpha) = n$, and where $\mathfrak{R}$ is an algebrically closed formally complex field of infinite transcendence over the rationals. Thus, $\mathfrak{R}$ may be taken as the field of complex numbers. Most of the arguments given will also hold for any field without characteristic.

Results will mainly be concerned with the projective (n-1)-space $\mathfrak{P}$ whose points are the equivalence classes of proportional elements of $\mathscr{C}_{\alpha^*}$, other than zero.

B.1 Some general theorems.

In this chapter elementary concepts of invariant theory will be incorporated into a tensor scheme and proofs of results equivalent to Gram's theorem and the first fundamental theorem of invariant theory will be given.

B 1.1 Coordinates and forms.

Let $\mathscr{F}$ be a frame for $\mathscr{C}$ and let x be the point corresponding to an element x^α of $\mathscr{C}_{\alpha^*}$. (As with the more general case of α-Grassmann and dual α-Grassmann tensors, I use the same symbol for a point or prime of $\mathfrak{P}$ as for a tensor representing it, see A 4.4e) If $\mathbf{v}_{(1)\alpha}, \ldots, \mathbf{v}_{(n)\alpha}$, $\mathbf{w}^\alpha_{(1)}, \ldots, \mathbf{w}^\alpha_{(n)}$ are the elements of $\mathscr{F}$, then the coordinates in $\mathscr{F}$ of x

are $v_{(1)} \cdot x$, .., $v_{(m)} \cdot x$. Similarly, if y is a prime in $\mathfrak{B}$ corresponding to $y_\alpha \in \mathcal{C}$, then $y \cdot w_{(1)}$, .., $y \cdot w_{(m)}$ is a set of coordinates of the prime y.

B 1.1a Homogeneous polynomials.

Let x^1, .., x^m ; .... ; z^1, .., z^m be coordinates in $\mathfrak{F}$, of the points x, ..., z of $\mathfrak{B}$, respectively. Let $a(x^1, \ldots, x^m; \ldots ; z^1, \ldots, z^m)$ be a form over $\mathfrak{R}$ in the sets of variables $x^1, \ldots, x^m; \ldots ; z^1, \ldots, z^m$. Then $a(x^1, \ldots, x^m; \ldots ; z^1, \ldots, z^m)$

$$= \sum a_{i_1 \ldots i_\lambda, \ldots, k_1 \ldots k_t} \; x^{i_1} \ldots x^{i_\lambda} \ldots z^{k_1} \ldots z^{k_t}$$

summed over all sets $(i_1, \ldots, k_t)$ where $i_1, \ldots, k_t$ take values from 1, 2, .., n and where the $a_{i_1 \ldots \ldots k_t}$ are elements of $\mathfrak{R}$. Now $x^i = v_{(i)} \cdot x$; ... ; $z^i = v_{(i)} \cdot z$, so that putting

$$a = \sum (a_{i_1 \ldots i_\lambda, \ldots, k_1 \ldots k_t} \; v_{(i_1)} \ldots v_{(i_\lambda)} \ldots v_{(k_1)} \ldots v_{(k_\lambda)}), \text{ we get}$$

$$a(x^1, \ldots, x^m; \ldots ; z^1, \ldots, z^m) = a \cdot \{(x)^\lambda \ldots (z)^t\} .$$

Also, from A 3.1g we see that $a_{i_1 \ldots i_\lambda, \ldots, k_1 \ldots k_t}$ are the components of a in the frame $\mathfrak{F}$, thus justifying the notation.

This shows that any form in the variables corresponds to a tensor polynomial of the type given above.

More generally, we have the following result :

B 1.1b Lemma. Any form in the components, in some frame $\mathfrak{F}$, of the tensors x^β, ..., z^δ of $\mathcal{C}$, can be written in the form

$$a \cdot \{(x)^\lambda \ldots (z)^t\}$$

where $a_{\beta^\lambda \ldots \delta^t} \in \mathcal{C}$ and r, .., t are the respective degrees of the

polynomial in the components of x, ..., z.

B 1.1c In accordance with the usual nomenclature, an element of $\mathcal{C}_{\alpha^\wedge}$ will sometimes be called a <u>covariant</u> tensor and an element of $\mathcal{C}_{\alpha^{*\wedge}}$ a <u>contravariant</u> tensor. An element of $\mathcal{C}_{\alpha^\wedge\,\alpha^{*\wedge}}$ may be called a <u>mixed</u> tensor ($rs > 0$).

B 1.2 <u>Theorem</u>. Let $\beta, .., \delta \in \mathfrak{S}$ and let $a^{\epsilon}_{\beta^\wedge..\delta^t} \in \mathcal{C}$; then $a\bullet\{(x)^\wedge .. (z)^t\} = 0$ for all $x^\beta \in \mathcal{C}_{\beta^*}$, .., $z^\delta \in \mathcal{C}_{\delta^*}$, if and only if

$$a\bullet\{P^{(\wedge)}_{(\beta)} .. P^{(t)}_{(\delta)}\} = 0.$$

<u>Proof</u>. Consider, first, the case of one variable only, that is, $a\bullet(x)^\wedge = 0$ for all $x^\beta \in \mathcal{C}_{\beta^*}$, where $a^{\epsilon}_{\beta^\wedge} \in \mathcal{C}_{\beta^\wedge\epsilon^*}$. Let $\mathcal{F}$ be a frame in $\mathcal{C}$. The components in $\mathcal{F}$ of the tensor $a\bullet(x)^\wedge$ are then homogeneous polynomials in the components of x as in B 1.1a. Thus $a\bullet(x)^\wedge = 0$ for all x^β if and only if all the coefficients of these polynomials are zero. Now each coefficient is the sum of those components which correspond to the same product of components of x. It follows that the tensor formed by symmetrizing the lower β-indices of a vanishes, that is, $a\bullet P^{(\wedge)}_{(\beta)} = 0$. The result follows by induction.

B 1.2a <u>Corollary</u>. The function $a\bullet\{(x)^\wedge .. (z)^t\}$ of the variables x^β, .., z^δ uniquely determines the tensor $b = a\bullet\{P^{(\wedge)}_{(\beta)} .. P^{(t)}_{(\delta)}\}$. Also, $a\bullet\{(x)^\wedge .. (z)^t\} = b\bullet\{(x)^\wedge .. (z)^t\}$.

<u>Proof</u>. Suppose $a\bullet\{(x)^\wedge .. (z)^t\} = a'\bullet\{(x)^\wedge .. (z)^t\}$ for all

$x^\beta \in \mathcal{C}_{\beta^*}, \ldots, z^\delta \in \mathcal{C}_{\delta^*}$; then, by theorem B 1.2 $(a - a') \cdot \{P^{(\lambda)}_{(\beta)} \ldots P^{(\tau)}_{(\delta)}\} = 0$ so that $b = a' \cdot \{P^{(\lambda)}_{(\beta)} \ldots P^{(\tau)}_{(\delta)}\}$ also. That $a \cdot \{(x)^\lambda \ldots (z)^\tau\} = b \cdot \{(x)^\lambda \ldots (z)^\tau\}$ follows from the relation

$$P^{(\lambda)}_{(\beta)} \cdot (x)^\lambda = (x)^\lambda \qquad \text{B 1.2b}$$

for each $x^\beta \in \mathcal{C}_{\beta^*}$.

B 1.2c It will often be convenient, when considering a form $a(x^1, \ldots, x^n ; \ldots ; z^1, \ldots, z^n) = a \cdot \{(x)^\lambda \ldots (z)^\tau\}$, to study instead the tensor $b = a \cdot \{P^{(\lambda)}_{(\beta)} \ldots P^{(\tau)}_{(\delta)}\}$, since, by B 1.2a, b is uniquely determined by the form. I shall be concerned, to a large extent, with forms which are not, in fact, general forms. That is, forms whose coefficients satisfy a set of algebraic relations which are independent of the frame of reference. Now, such a condition is expressed, in invariant theory, by the vanishing of a set of covariants. Theorem B 1.3 gives a tensor equivalent of this result.

Some preliminary lemmas will first be proved.

B 1.2d Lemma. Let $b^\beta \in \mathcal{C}$ and $\mathcal{F}$ be a frame for $\mathcal{C}$. Then there exist integers k and l and elements $B^{\beta\, \alpha^m k}_{(i_1, \ldots, i_m)\, \alpha^l}$ of $\mathcal{J}$ such that

$$b = I_{(\beta)} (\epsilon^*)^k \cdot \left\{ \sum_{i_1 \ldots i_m} B_{(i_1, \ldots, i_m)} \cdot (w_{(1)})^{i_1} \ldots (w_{(m)})^{i_m} \right\}$$

where $w^\alpha_{(1)}, \ldots, w^\alpha_{(m)} \in \mathcal{F}$.

If b has rational components in $\mathcal{F}$, then the B's may be chosen to be elements of $\mathcal{R}$.

Proof. Put $\beta = \alpha^m \alpha^{*k}$ and $c^{\alpha^m}_{\alpha^k} = b^\beta$. Then, by A 4.5h, we have

$$c = \frac{(-1)^{k(m-k)}}{\{(m-1)!\}^k} c\left\{(\epsilon^*)^k \cdot D^{(m-1:1;k)*}\right\} \cdot \left\{D^{(1:m-1;k)} \cdot (\epsilon)^k\right\} I^{(k)}\Big\}.$$

Thus c is given by

$$c = I^{(m)}\left\{(\epsilon^*)^k \cdot D^{(m-1:1;k)}\right\} \cdot dI^{(k)}$$

where $d = d^{\alpha^l} \in \mathcal{C}$, $l = m + nk - k$ and d has rational components in if d has. (The tensor ϵ is chosen to have components ± 1 or 0, see A 4.5a.) Let $d^{j_1 \cdots j_l}$ be the components of d in $\mathcal{F}$. Then, by A 3.1i we have $d = \sum d^{j_1 \cdots j_l} w_{(j_1)} \cdots w_{(j_l)} = \sum_{i_1+\cdots+i_m=l} \sum^{l!} \lambda_{\lambda; i_1, \ldots, i_m} J_\lambda \cdot (w_{(1)})^{i_1} \cdots (w_{(m)})^{i_m}$ where the λ's are rational if the components of d are rational. Thus $d = \sum_{i_1+\cdots+i_m=l} C_{(i_1, \ldots, i_m)} \cdot (w_{(1)})^{i_1} \cdots (w_{(m)})^{i_m}$ where the C's are elements of $\mathcal{F}$ and are elements of $\mathcal{R}$ if b has rational components. The result follows.

B 1.2e **Lemma.** Let $\beta \in \mathfrak{S}$ and let the tensor $a_\beta \in \mathcal{C}$ satisfy a condition given by a number of linear relations, homogeneous in the components of a in a given frame $\mathcal{F}$, and with [rational] coefficients in $\mathcal{R}$. If the condition on a, thus obtained, is independent of $\mathcal{F}$, then it may be written $a \cdot F = 0$ where F^β_β is a projector in $\mathcal{F}$ [in $\mathcal{R}$].

If the linear relations are written as $a \cdot b_{(j)} = 0$; $j \in \Omega$, where Ω is some set of suffixes, then $F \cdot b_{(j)} = b_{(j)}$.

Proof. By B 1.1b, the relations can always be written in the form $x \cdot b_{(i)} = 0$; $i \in \Omega$ [where the $b_{(i)}$ have rational components in $\mathcal{F}$].

Also from B 1.2d we have $b_{(j)} = I_{(\beta)} (\epsilon^*)^k \cdot \left\{\sum_{i_1+\cdots+i_m=l} B_{(j; i_1, \ldots, i_m)} \cdot (w_{(1)})^{i_1} \cdots (w_{(m)})^{i_m}\right\}$;

$j \in \Omega$, where the B's are elements of $\mathfrak{F}$ [of $\mathfrak{R}$].

Now, if the relations $x \cdot b_{(j)}$; $j \in \Omega$ are to represent a condition on x which is independent of $\mathfrak{F}$, they must also hold if each $w_{(i)}$ is replaced by $\mu_i w_{(i)}$ where $0 \neq \mu_i \in \mathfrak{K}$, since a new frame could be chosen for with $\mu_i w_{(i)}^{\alpha}$ for elements. Thus, without loss of generality, we can suppose that for each j, the $B_{(j;i_1,\ldots,i_m)}$ are non-zero for only one set of $i_1, \ldots, i_m$ (i.e. the relations are isobaric) so that the tensors $b_{(j)}$ are given by

$$b_{(j)} = I_{(\beta)} (\epsilon^*)^{k} \cdot B_{(j)} \cdot (w_{(1)})^{i_{1j}} \ldots (w_{(m)})^{i_{mj}} \quad ; \quad j \in \Omega$$

where the $B_{(j)}$ are some elements of $\mathfrak{F}$ [of $\mathfrak{R}$] and the i_{ij} are integers, positive or zero, such that $i_{1j} + \ldots + i_{mj} = 1$. Now the conditions $x \cdot b_{(j)} = 0$; $j \in \Omega$ must hold independently of $\mathfrak{F}$, that is, independently of whatever set of linearly independent elements $w_{(1)}, \ldots, w_{(m)}$ of $\mathfrak{C}_{\alpha}^*$ are chosen (see A 2.7c). However, the condition that $w_{(1)}, \ldots, w_{(m)}$ be linearly independent is clearly irrelevant ($\mathfrak{K}$ is infinite, so that the principle of irrelevance of algebraic inequalities applies). Thus, the relations $\{a(\epsilon^*)^{k}\} \cdot B_{(j)} \cdot \{(w_{(1)})^{i_{1j}} \ldots (w_{(m)})^{i_{mj}}\} = 0$ must hold identically in $w_{(1)}, \ldots, w_{(m)}$, since all α^*-alternating tensors are proportional, so that by B 1.2, the condition on a becomes $a(\epsilon^*)^{k} \cdot A_{(j)} = 0$; $j \in \Omega$, where for each j, $A_{(j)} = B_{(j)} \cdot \{P^{(l_{j1})} \ldots P^{(l_{jm})}\}$. Now $P^{(\lambda)} \in \mathfrak{R}$ so that $A_{(j)} \in \mathfrak{F}$ [$A_{(j)} \in \mathfrak{R}$]. The relations $a(\epsilon^*)^{k} \cdot A_{(j)} = 0$ are equivalent to $a(N^{(m)})^{k} \cdot A_{(j)} = 0$; $j \in \Omega$, by A 4.5b (since $\epsilon \neq 0$). Putting

$C^{\beta}_{\alpha^{*}\ldots\alpha'} = I_{(\beta)}(N^{(\omega)})^{k}\cdot A_{(j)}$, the condition becomes $a\cdot C_{(j)} = 0$; $j \in \Omega$.
Let $F = [C_{(j)}\cdot C^{\dagger}_{(j)}$; $j \in \Omega]$; then (by A4.31, A1.6p) F is a projector in $\mathcal{F}$ [in $\mathcal{R}$] since the $C_{(j)}$ are in $\mathcal{F}$ [in $\mathcal{R}$]. Also, by A 4.1c and A 4.2c, we see that those tensors of $\mathcal{C}_{\beta}$ which satisfy all the relations $a\cdot C_{(j)} = 0$ are just those satisfying $a\cdot F = 0$. But the tensors a_{β} which satisfy all of $a\cdot C_{(j)} = 0$ are those satisfying the original condition given by $a\cdot b_{(j)} = 0$, $j \in \Omega$. Thus, $F = [b_{(j)}\, b^{\dagger}_{(j)}$; $j \in \Omega]$ also. We have $F\cdot b_{(i)}$
$= F\cdot(b_{(i)}\, b^{\dagger}_{(i)})\cdot b_{(i)} = b_{(i)}\, b^{\dagger}_{(i)}\cdot b_{(i)} = b_{(i)}$, by A 4.2d . This proves the result.

B 1.2f Remark. In the proof of B 1.2e, the device is used of treating the contravariant indices of a with the aid of an alternating tensor. This enables us to treat mixed tensors in terms of covariant tensors. However, if in B 1.2e, the tensor a had been simply a covariant (or contravariant) tensor, it would have been possible to omit the word "homogeneous" from the conditions of the lemma. This cannot be done in the general case because of the possibility of non-homogeneous conditions such as $a^{\alpha}_{\alpha} = I_{(\omega)}$.

Lemma B 1.2e is closely connected with the fact that the only tensors invariant under all linear transformations are the elements of $\mathcal{F}$.
In fact, the proof of B 1.2 e is essentially a demonstration that the only projectors invariant under all linear transformations are elements of $\mathcal{F}$.

B 1.2g Lemma. Let $\beta, \ldots, \delta \in \mathfrak{S}$, let $\zeta = \beta^{\wedge}\ldots\delta^{t}$ and let the tensors

a_β , .., $c_\delta \in \mathcal{C}$ satisfy the condition $\{(a)^\lambda .. (c)^\tau\} \cdot f = 0$, where f^δ is an element of $\mathcal{C}$. Suppose that a, .., c also satisfy any condition $\{(a)^\lambda .. (c)^\tau\} \cdot g = 0$ where g is the image of f under some linear transformation. Then there exists a projector F^δ_ζ of $\mathcal{F}$ such that $\{(a)^\lambda .. (c)^\tau\} \cdot F = 0$ and $F \cdot f = f$.

Proof. If g is the image of f under the linear transformation Φ, then the components of g in $\Phi(\mathcal{F})$ are just those of f in $\mathcal{F}$. Thus, replacing the tensor a, of B 1.2e, by $(a)^\lambda .. (c)^\tau$ and the tensors $b_{(i)}$ by f and the images of f under the various linear transformations and using B 1.2e, we obtain the projector F as required.

B 1.3 Theorem. Let β , .., $\delta \in \mathfrak{S}$ and let the tensors a_β , .., $c_\delta \in \mathcal{C}$ satisfy a condition given by the vanishing of a number of polynomials over the rationals [over $\mathfrak{R}$], which are homogeneous in the components in a frame $\mathcal{F}$, of a, ..., c, separately. If the condition on a,.., c thus obtained, is independent of $\mathcal{F}$, then it is equivalent to the condition

$$\{(a)^\lambda .. (c)^\tau\} \cdot F = 0,$$

where F is some projector in $\mathcal{R}$ [in $\mathcal{F}$] and where r, .., t are integers. (If the number of polynomials is finite, r, ..., t can be taken as the greatest of the respective degrees to which the components of a, .., c appear.)

Conversely, any condition of the form $\{(a)^\lambda .. (c)^\tau\} \cdot F = 0$ where $F \in \mathcal{F}$ is the same condition on the components of a, .., c in any

frame of $\mathcal{C}$.

Proof. By the Hilbert basis theorem and by B 1.1b, the polynomials are of the form

$$\{(a)^{\lambda_i} \,..\, (c)^{t_i}\} \cdot f_{(i)} = 0 \qquad (i = 1, .., m).$$

Let $r = \max(r_1, .., r_m)$, .., $t = \max(t_1, .., t_m)$ and put $g_{(i)} = f_{(i)} I_{(\beta^{\lambda-\lambda_i} .. \delta^{t-t_i})}$. The above relations are then equivalent to

$$\{(a)^{\lambda_i} \,..\, (c)^{t_i} (a)^{\lambda-\lambda_i} .. (c)^{t-t_i}\} \cdot g_{(i)} = 0 \qquad (i = 1, .., m);$$

that is, to

$$\{(a)^{\lambda} \,..\, (c)^{t}\} \cdot h_{(i)} = 0 \qquad (i = 1, .., m)$$

where $h_{(i)}$ is a suitable transpose of $g_{(i)}$ and thus has rational components [components in $\mathfrak{R}$] in the frame $\mathcal{F}$. Now, these equations can be regarded as a set of homogeneous linear conditions on the components of the tensor $(a)^{\lambda} .. (c)^{t}$ in the frame $\mathcal{F}$. Thus, by B 1.2e, the condition on a, .., c is equivalent to $\{(a)^{\lambda} .. (c)^{t}\} \cdot F = 0$, where F is some projector in $\mathcal{R}$ [in $\mathcal{F}$]. The converse result follows from A 3.3n.

B 1.3a Elementary conditions. The connection between theorem B 1.3 and the classical form for invariant equations will now be discussed (see also Littlewood (12)).

Let the tensors a_β, .., c_δ of theorem B 1.3 be covariant and let the integer p be defined by $\alpha^p = \beta^{\lambda} .. \delta^{t}$. For each partition λ of p, the expressions $T^{\{\lambda\}}$, $J^{\{\lambda\}}$, $K^{\{\lambda\}}$ and $G^{\{\lambda\}}$ are defined as in A 1.8b, A 1.8g, A 1.8h, A 1.8j. By A 1.8c and A 1.8q we have the following

two relations

$$I = \sum_{\lambda} T^{\{\lambda\}} \quad \text{and} \quad T^{\{\lambda\}} = \frac{\chi_i^{(\lambda)}}{p!} \sum_{i=1}^{p!} J_i \cdot G^{\{\lambda\}} \cdot J_i^* .$$

$$I = \frac{1}{p!} \sum_{\lambda} \sum_{i=1}^{p!} \chi_i^{(\lambda)} J_i \cdot G^{\{\lambda\}} \cdot J_i^* .$$

Now let $d = (a)^{\alpha} \,..\, (c)^{\epsilon}$ and let $F \in \mathfrak{F}_{(\alpha)}$. We have

$$d \cdot F = \frac{1}{p!} \sum_{\lambda} \sum_{i=1}^{p!} \chi_i^{(\lambda)} d \cdot F \cdot J_i \cdot G^{\{\lambda\}} \cdot J_i^* .$$

Hence, the relation $d \cdot F = 0$ is equivalent to the set of relations $d \cdot F \cdot J_i \cdot G^{\{\lambda\}} \cdot J_i^* = 0$, i.e. to $d \cdot F \cdot J_i \cdot K^{\{\lambda\}} = 0$ (by A 1.8j and A 1.6n) where $i = 1, .., p!$ and λ ranges through the partitions of p. Expressing F as a linear combination of the permutations J_i, we see that the relation $d \cdot F = 0$ is equivalent to the vanishing of certain linear combinations (with rational coefficients, if $F \in \mathfrak{R}$) of the elementary expressions $d \cdot J_i \cdot K^{\{\lambda\}}$, for $i = 1, .., p!$ and for all partitions λ of p. (They are not all linearly independent.)

The tensor $d \cdot J_i \cdot K^{\{\lambda\}}$ is a <u>Clebsch tensor of type $\{\lambda\}$</u>, and $d \cdot J_i \cdot K^{\{\lambda\}} = 0$ represents an <u>elementary</u> condition on a, .., c.

B 1.3b <u>Clebsch forms</u>. Let the tensors $z_{(1)}^{\alpha}$, $z_{(2)}^{\alpha}$, ..., $z_{(i)}^{\alpha}$, .. be variable elements of $\mathfrak{C}_{\alpha^*}$ and define the tensors $x^{(1)}$, $x^{(2)}$, ..., $x^{(i)}$, .. by $x^{(1)} = z_{(1)}$, $x^{(2)} = N^{(2)} \cdot (z_{(1)} \; z_{(2)})$, ..., $x^{(i)} = N^{(i)} \cdot (z_{(1)} \,..\, z_{(i)})$, ... (that is, $x^{(i)\alpha_1 \ldots \alpha_i} = \frac{1}{i!} \begin{vmatrix} z_{(1)}^{\alpha_1} & .. & z_{(1)}^{\alpha_i} \\ \vdots & & \vdots \\ z_{(i)}^{\alpha_1} & .. & z_{(i)}^{\alpha_i} \end{vmatrix}$.) Note that $x^{(i)} = 0$ if $i > n$.

A <u>Clebsch variable tensor of type $\{\lambda\}$</u> is then defined to be a tensor $v^{\{\lambda\}}$ of the form $v^{\{\lambda\}} = x^{(\mu_1)} \, x^{(\mu_2)} \,..\, x^{(\mu_s)}$

where $(\mu_1 \mu_2 .. \mu_q)$ is the partition of p conjugate to $\lambda = (\lambda_1 \lambda_2 .. \lambda_{q'})$ (see Littlewood (14) p.287). We have,

$$v^{\{\lambda\}} = N^{(\mu_1)} N^{(\mu_2)} \ldots N^{(\mu_q)} \cdot J^{\{\lambda\}*} \cdot (z_{(1)})^{\lambda_1} (z_{(2)})^{\lambda_2} \ldots (z_{(q')})^{\lambda_{q'}}$$

A <u>Clebsch form of type $\{\lambda\}$</u> is an expression of the form $f \cdot v^{\{\lambda\}}$ where $v^{\{\lambda\}}$ is a Clebsch variable tensor of type $\{\lambda\}$. It follows from B 1.2, B 1.2a and the above expression for $v^{\{\lambda\}}$ that this form determines the tensor $f \cdot K^{\{\lambda\}}$ uniquely and that $f \cdot v^{\{\lambda\}}$ vanishes identically if and only if $f \cdot K^{\{\lambda\}} = 0$. The form $\{(a)^{\alpha} .. (c)^{\epsilon}\} \cdot A \cdot v^{\{\lambda\}}$, where $A \in \mathcal{F}$, is called a <u>covariant of type $\{\lambda\}$</u> of the set of covariant tensors a, .., c. Thus, this covariant vanishes identically if and only if $\{(a)^{\alpha} .. (c)^{\epsilon}\} \cdot A \cdot K^{\{\lambda\}} = 0$.

The combination of the results of B 1.3a and B 1.3b tells us that if the covariant tensors a_β, .., c_δ satisfy an invariant relation $\{(a)^{\alpha} .. (c)^{\epsilon}\} \cdot F = 0$, then this is equivalent to the identical vanishing of all the Clebsch forms

$$\{(a)^{\alpha} .. (c)^{\epsilon}\} \cdot F \cdot J_{(\lambda)} \cdot v^{\{\lambda\}} .$$

This result, or rather the result together with B 1.4 to follow, gives a proof of Gram's theorem on invariant equations (see Turnbull (21) p.271).

B 1.3c <u>Invariants</u>. It will be observed that the non-zero tensors $x^{(1)}$, $x^{(2)}$, .., $x^{(i)}$, .. are α-Grassmann tensors. In particular, $x^{(m)}$ is an α-alternating tensor. Now, all such tensors are proportional, so

that, without loss of generality, we can put $x^{(\alpha)} = \epsilon^{(\alpha)}$.

The invariants of the set of covariant tensors a_β, .., c_δ are Clebsch forms of type $\{q^n\}$, that is an expression of the form

$$\{(a)^\lambda \,..\, (c)^\tau\} \cdot A \cdot (\epsilon^{(\alpha)})^\rho,$$

where $A^\eta_\eta \in \mathfrak{F}$.

However, I shall prefer to call such expressions <u>simple invariants</u> (for reasons which will appear later, see B4.7j). More generally, I define a simple invariant of the tensors a_β, .., c_δ, where a, .., c need not necessarily just be covariant, to be an expression of the form

$$\Psi = \{(a)^\lambda \,..\, (c)^\tau\} \cdot A \cdot \{(\epsilon^{(\alpha)})^{\rho_1} (\epsilon^{(\alpha^*)})^{\rho_2}\},$$

where $A \in \mathfrak{F}$. If A is proportional to an element of $\mathfrak{R}$, Ψ will be called a simple invariant over $\mathfrak{R}_1$.

The <u>weight</u> of Ψ is defined to be $\rho_1 - \rho_2$. (By A 4.5b, $\epsilon^{(\alpha)} \epsilon^{(\alpha^*)}$ is proportional to $N^{(\alpha)}$ ($\in \mathfrak{R}$) and it is in fact unnecessary to have both of ρ_1 and ρ_2 greater than zero in the definition of the simple invariant Ψ. In any particular case, only one of the tensors $\epsilon^{(\alpha)}$, $\epsilon^{(\alpha^*)}$ need occur.)

A <u>simple absolute invariant</u> of a, .., c is defined to be the quotient of two simple (relative) invariants of equal weight.

The following theorem is essentially Littlewood's fundamental theorem for concomitants of a set of tensors (see (12)). It is also essentially the first main theorem of invariant theory (see B 1.4a).

B 1.4 **Theorem.** Let $\mathfrak{Q}$ be a class of sets $(a_\beta, .., c_\delta)$ of elements of $\mathfrak{C}$. Let $\phi = \phi(a, .., c)$ be a polynomial over $\mathfrak{R}$ [over the rationals] homogeneous in the components of a, .., c separately, in some frame $\mathfrak{F}$, and suppose that the vanishing of ϕ represents the same condition on the sets (a, .., c) of $\mathfrak{Q}$ irrespectively of the choice of $\mathfrak{F}$. Then there is a simple invariant $\psi(a, .., c)$ of a, .., c [over $\mathfrak{R}_1$] such that $\psi(a, .., c) = \phi(a, .., c)$, for all $(a, .., c) \in \mathfrak{Q}$.

Proof. We can write $\phi(a, .., c) \equiv (a)^k .. (c)^m \cdot d$, where $d^\eta \in \mathfrak{C}_{\eta^*}$, $\eta = \beta^k .. \delta^m$ and k, .., m are the degrees to which the components of a, .., c respectively appear (see B 1.1b). By B 1.2d we can put $d = I_{(\eta)} (\epsilon^*)^\wedge \cdot \left\{ \sum_{i_1 + .. + i_n = p} B_{(i_1, .., i_n)} \cdot (z_{(1)})^{i_1} .. (z_{(m)})^{i_m} \right\}$ where $z^\alpha_{(1)}, .., z^\alpha_{(m)} \in \mathfrak{F}$ and where the B's are elements of $\mathfrak{P}$ [of $\mathfrak{R}$]. Now suppose that $\phi(a, .. c) \neq 0$ for some $(a, .., c) \in \mathfrak{Q}$. (Otherwise the result is trivial, with $\psi \equiv 0$.) Then the above sum has just one non-vanishing term (so that ϕ is isobaric) since otherwise, replacing $\mathfrak{F}$ by a frame $\mathfrak{F}'$ where $\mu_i z_{(i)} \in \mathfrak{F}'$ $(i = 1, .., n)$, $\mu_i \neq 0$, we can choose the μ_i's so that the corresponding polynomial $\phi'(a, .., c)$ $= (a)^k .. (c)^m (\epsilon^*)^\wedge \cdot \left\{ \sum \mu_1^{i_1} .. \mu_m^{i_m} B_{(i_1, .., i_m)} \cdot (z_{(1)})^{i_1} .. (z_{(m)})^{i_m} \right\}$ does not vanish but $\phi(a, .., c)$ does. Thus we have $\phi(a, .., c)$ $\equiv (a)^k .. (c)^m (\epsilon^*)^\wedge \cdot B \cdot (z_{(1)})^{i_1} .. (z_{(m)})^{i_m}$ for some $i_1, .., i_m$ and some $B \in \mathfrak{P}$ $[B \in \mathfrak{R}]$.

Now, the condition of the theorem is that for each frame $\mathfrak{F}$, that is

for each set of linearly independent elements $z_{(1)}, .., z_{(m)}$ of $\mathcal{C}_{\alpha^*}$, this expression vanishes for the same sets of tensors a, .., c for which $(a, .., c) \in \mathfrak{Q}$. Choose any fixed set (a, .., c) of $\mathfrak{Q}$ and consider $\{(a)^{\lambda} .. (c)^{\mu} (\epsilon^*)^{\nu} \bullet B\} \bullet (z_{(1)})^{i_1} .. (z_{(m)})^{i_m} = g \bullet (z_{(1)})^{i_1} .. (z_{(m)})^{i_m}$ as a function of $z_{(1)}, .., z_{(m)}$. Suppose first that this function is not identically zero in $z_{(1)}, .., z_{(m)}$. Then it is a function which can vanish only when $z_{(1)}, .., z_{(m)}$ are linearly dependent.

Now, each $z_{(i)}$ represents a point in $\mathfrak{B}$, and the set of points $(z_{(1)}, .., z_{(m)})$ can be represented by a point on a Segre variety Ω. The relation $g \bullet (z_{(1)})^{i_1} .. (z_{(m)})^{i_m} = 0$ corresponds to a hypersurface $\prod$ on Ω. But $g \bullet (z_{(1)})^{i_1} .. (z_{(m)})^{i_m}$ can vanish only when $z_{(1)}, .., z_{(m)}$ are linearly dependent. Thus, $\prod$ consists of components of the subvariety Λ of Ω representing the sets of linearly dependent $z_{(1)}, .., z_{(m)}$. But Λ is irreducible (since all sets of n linearly independent points of are on an equal footing) and is the prime section of Ω given by $\epsilon^* \bullet z_{(1)} .. z_{(m)} = 0$ (see A 2.7 and A 4.5b). Hence

$$g \bullet (z_{(1)})^{i_1} .. (z_{(m)})^{i_m} \equiv \zeta \{\epsilon^* \bullet z_{(1)} .. z_{(m)}\}^q$$

for some integer q and some scalar ζ (independent of $z_{(1)} .. z_{(m)}$). Thus $i_1 = .. = i_m = q$ and

$$g \bullet (P^{(q)})^{\kappa} = \zeta (\epsilon^*)^q \bullet J \bullet (P^{(q)})^{\kappa}$$

for a suitable permutation J (by B 1.2).

This will also hold if $g \bullet (z_{(1)})^{i_1} .. (z_{(m)})^{i_m} \equiv 0$, with $\zeta = 0$,

(since $\phi(a, .., c)$ is sometimes non-zero for $(a, .., c) \in \mathfrak{Q}$ so that $i_1, .., i_n$ are fixed). Now $\epsilon^* = \epsilon^* \bullet N^{(n)}$ and $(N^{(n)})^q \bullet J \bullet (P^{(q)})^n = K^{\{q^n\}} = K$ say.

Hence,

$$g \bullet (P^{(q)})^n = \xi (\epsilon^*)^q \bullet K = \kappa \xi (\epsilon^*)^q \bullet K \bullet K^* \bullet K = \kappa g \bullet (P^{(q)})^n \bullet K^* \bullet K$$

$$= \kappa\, g \bullet K^* \bullet (N^{(n)})^q \bullet K = \frac{\kappa}{(n!)^q} \{d \bullet K^* \bullet (\epsilon)^q\} \{ (\epsilon^*)^q \bullet J \bullet (P^{(q)})^n \}$$

where $\kappa = \kappa_{(q^n)} \left(= \frac{(q!)^n (n!)^q}{(qn)!} \chi_1^{(q^n)}\right)$, by A 1.8i and A 4.5b.

Thus,

$$d \bullet (z_{(1)})^q \,..\, (z_{(n)})^q = \left\{\frac{\kappa}{(n!)^q} (\epsilon^* \bullet z_{(1)} \,..\, z_{(n)})^q\right\} d \bullet K^* \bullet (\epsilon)^q .$$

But $\xi = \frac{\kappa}{(n!)^q} (\epsilon^* \bullet z_{(1)} \,..\, z_{(n)})^q$ is independent of a, .., c. Hence, putting $\psi(a, .., c) = \xi\ (a)^\lambda \,..\, (c)^m (\epsilon^*)^\lambda \bullet B \bullet K^* \bullet (\epsilon)^q$, we get $\psi(a, .., c) = \phi(a, .., c)$ for all $(a, .., c) \in \mathfrak{Q}$, and ψ is a simple invariant of a, .., c as required.

B 1.4a Let $\phi(a, .., c)$ be an invariant of a, .., c in the classical sense, that is, if a′, .., c′ are the images of a, .., c under some linear transformation Θ of $\mathfrak{C}$, then $\phi(a', .., c') = \theta\ \phi(a, .., c)$ where $\theta (\neq 0)$ is some scalar depending only on Θ. It is clear that such a function ϕ satisfies the conditions of theorem B 1.4 (with $\mathfrak{Q} = \mathfrak{C}_\beta \times .. \times \mathfrak{C}_\delta$). Hence, by B 1.4, $\phi(a, .., c)$ is a simple invariant in the sense of B 1.3c, so that it is of the form

$$\phi(a, .., c) = \{(a)^\lambda \,..\, (c)^m\} \bullet B \bullet (\epsilon^{(\alpha)})^q (\epsilon^{(\alpha^*)})^{q'}$$

The $\epsilon^{(\alpha)}$ and $\epsilon^{(\alpha^*)}$ give rise to the bracket factors of the classical

invariant formalism. Thus, the first fundamental theorem of invariant theory is a consequence of B 1.4, the theorem for covariants being obtained by considering the simultaneous system consisting of the given tensors together with the variable tensors (see Turnbull (21) pp. 207). Theorem B 1.4 is, in fact, somewhat more general, since the class $\mathfrak{Q}$ is unrestricted.

B.1.5 Projections of a tensor system.

Let e_{α}^{α} be any idempotent element of $\mathcal{C}_{\alpha\alpha^*}$. Then e determines a mapping Ψ of the elements of $\mathcal{C}$ onto a subset $\mathcal{C}'$ of $\mathcal{C}$ in the following way.

If $a_{\alpha_1 \cdots \alpha_s}^{\alpha_1 \cdots \alpha_r}$ is any element of $\mathcal{C}$, then the image a′ of a under Ψ is defined by

$$a' = \Psi(a) = (e)^{r} \cdot a \cdot (e)^{s}.$$

Thus, the set $\mathcal{C}'$ consists of those elements $x_{\alpha^t}^{\alpha^q}$ of $\mathcal{C}$ which satisfy

$$x = (e)^{q} \cdot x \cdot (e)^{t}$$

(that is, $x = \Psi(x)$), since e is idempotent.

B 1.5a If a and b are two elements of $\mathcal{C}$, then $\Psi(a+b) = \Psi(a) + \Psi(b)$ and $\Psi(ab) = \Psi(a)\Psi(b)$. Also, if $J_i \cdot a \cdot J_j$ is any transpose of a then the corresponding transpose of $\Psi(a)$ is given by $J_i \cdot \Psi(a) \cdot J_j = \Psi(J_i \cdot a \cdot J_j)$. These properties follow immediately from the definition of Ψ.

If a is in $\mathcal{C}'$, then clearly any contraction of a is also in $\mathcal{C}'$.

Hence $\mathcal{C}'$ is a tensor system and the tensor operations of $\mathcal{C}'$ are given by those of $\mathcal{C}$.

B 1.5b The α-unit of the system $\mathcal{C}'$ is e. The integer $m = \langle e \rangle$ will be called the rank of Ψ.

B 1.5c By A 4.4n we have $N^{(m)}\bullet(e)^m = \frac{1}{(q\bullet p)} pq$, where $q\bullet p \neq 0$ and where p^{α^m} and q_{α^m} are, respectively, a Grassmann tensor of an $(m-1)$-space in $\mathfrak{B}$ and a dual Grassmann tensor of an $(n-m-1)$-space in $\mathfrak{B}$ which does not meet the $(m-1)$-space p (see A 4.4g). The space p will be called the base space of Ψ and the space q, the vertex space of Ψ. Since, By A 4.4l, we have $e = \frac{(-1)^{\lambda-1}\lambda}{(q\bullet p)} I^{(1)} q\bullet p I^{(1)}$, it follows that any projection Ψ is determined by its base space and vertex space. Furthermore, any such pair of spaces p and q determine a projection of $\mathcal{C}$ $(q\bullet p \neq 0)$.

B 1.5d Applying A 4.4m to $N^{(m)}\bullet(e)^m = \frac{1}{(q\bullet p)} pq$ we see that $N^{(4)}\bullet(e)^4$ is proportional to $\frac{1}{(q\bullet p)} (pq)_{[\alpha^{m-4}]}$, i.e. to $I^{(4)} q\bullet p I^{(4)}$. Hence, by A 4.5d, $N^{(4)}\bullet(e)^4$ is proportional to $I^{(4)}(\epsilon^*\bullet I^{(n-m)} q')\bullet(p' I^{(m-m)}\bullet\epsilon) I^{(4)} = p' I^{(4)} \epsilon^* \bullet \epsilon I^{(4)} q'$ where $p'_{\alpha^{n-m}}$ and $q'^{\alpha^{n-m}}$ are, respectively, a dual Grassmann tensor corresponding to p and a Grassmann tensor corresponding to q.

Thus from A 4.5c we get

$$N^{(4)}\bullet(e)^4 = \mu\, p' I^{(4)} \bullet N^{(n-m+4)} \bullet q' I^{(4)}$$

for some $\mu \neq 0$.

B 1.5e Since $N^{(m)}\bullet(e)^m = \frac{1}{(q\bullet p)} pq$, the tensors p and q are elements of $\mathcal{C}'$ They are, in fact, an α-alternating tensor of $\mathcal{C}$ and an α^*-alternating

tensor of $\mathcal{C}'$, respectively (see A 4.5b). Thus, if a, .., c are elements of $\mathcal{C}'$ and $\phi(a, .., c)$ is a simple invariant of a, .., c in $\mathcal{C}'$, then ϕ is a simple invariant of weight zero, of a, .., c, p, q in $\mathcal{C}$ and conversely.

B 2 Non-general forms

A non-general form will be taken to mean a form whose coefficients satisfy a set of invariant algebraic relations. In this chapter, a technique for studying the invariant properties of such forms will be introduced.

B 2.1 Geometric and invariant varieties

Let the tensor x_β satisfy $(x)^\lambda \cdot F = 0$ where F is a projector in $\mathcal{P}$ The set $\mathcal{V}$ of all such tensors x will be called an invariant variety of $\mathcal{C}$, in the space $\mathcal{C}_\beta$. If $F \in \mathcal{R}$, $\mathcal{V}$ will be called a geometric variety. Clearly, any geometric variety is an invariant variety and any invariant variety is an algebraic affine variety. If x_β, .., z_ζ are elements of $\mathcal{C}$ satisfying $\{(x)^\lambda .. (z)^\tau\} \cdot F = 0$ where F is a projector in $\mathcal{R}$, then the set of such sets of tensors (x, .., z) constitute a geometric variety in the (affine) product space $\mathcal{C}_\beta \times .. \times \mathcal{C}_\zeta$ and similarly for invariant varieties in $\mathcal{C}_\beta \times .. \times \mathcal{C}_\zeta$.

B 2.1a Notation. The projective space associated with an affine space $\mathcal{C}_\beta$ will be denoted by $|\mathcal{C}_\beta|$. Similarly, the projective many-way space which is the product of a number of projective spaces $|\mathcal{C}_\beta|$, .., $|\mathcal{C}_\zeta|$ will be denoted by $|\mathcal{C}_\beta \times .. \times \mathcal{C}_\zeta| = |\mathcal{C}_\beta| \times .. \times |\mathcal{C}_\zeta|$. This notation can also

be extended to any subset $\mathfrak{Q}$ of $\mathfrak{C}_\beta \times .. \times \mathfrak{C}_\delta$. Thus $|\mathfrak{Q}|$ denotes the set of equivalence classes of sets $(a'_\beta , .., c'_\delta)$ where $a'_\beta , .., c'_\delta$ are proportional to $a_\beta, .., c_\delta$ respectively for some $(a, .., c) \in \mathfrak{Q}$ but where none of $a, .., c$ is zero.

B 2.1b Theorem B 1.3 tells us that the invariant varieties in $\mathfrak{C}_\beta$ are just those algebraic varieties in $\mathfrak{C}_\beta$ which are invariant under any linear transformation of $\mathfrak{C}$. Also, by A 3.3j, we see that the geometric varieties are invariant under any automorphism of $\mathfrak{C}$.

B 2.1c An invariant variety is <u>irreducible</u> if it is not the union of two invariant varieties unless one is contained in the other. Similarly, an irreducible geometric variety is one which is not the union of two geometric varieties unless one is contained in the other. If an invariant variety is reducible over $\mathfrak{R}$ when regarded as an algebraic variety, then it cannot be an irreducible invariant variety. For suppose $\mathfrak{V} = \mathfrak{V}_1 \cup \mathfrak{V}_2 \cup .. \cup \mathfrak{V}_\lambda$ expresses the invariant variety $\mathfrak{V}$ as a union of its component algebraic varieties, irreducible over $\mathfrak{R}$. Let Φ be any linear transformation of $\mathfrak{C}$; then $\Phi(\mathfrak{V}) = \mathfrak{V}$, so that for each i, $\Phi(\mathfrak{V}_i) = \mathfrak{V}_j$ for some j. Then we have a representation of the full linear group by the finite group of permutations of the $\mathfrak{V}_i$. That such a representation must be the identity, follows from the fact that any non-singular matrix over $\mathfrak{R}$ is the $(\lambda!)$-th power of another such matrix, since $\mathfrak{R}$ is algebraically closed and without characteristic (see Hodge and Pedoe (9)

pp(57-8). Thus $\Phi(\mathcal{V}_i^j) = \mathcal{V}_i^j$, for each linear transformation Φ of $\mathcal{C}$, so that each $\mathcal{V}_i^j$ is also an invariant variety.

The irreducible invariant varieties are, therefore, just those invariant varieties which are algebraic varieties which are irreducible over $\mathfrak{R}$. Similarly, the irreducible geometric varieties are just those geometric varieties which are algebraic varieties irreducible over the rationals $\mathcal{R}$, since each geometric variety is an invariant variety given by equations with rational coefficierts and conversely. It is possible for an irreducible geometric variety to be a reducible invariant variety (see B 2.5).

B 2.1d The intersection of any number or the join of a finite number of geometric [invariant] varieties is again a geometric [invariant] variety.

B 2.1e Definition. Let a_η be an element of $\mathcal{C}$ and define

$$E_{\{a\}} = I - [A \cdot A^\dagger \; ; \; a \cdot A = 0, \; A \in \mathcal{R}]$$

and $$E_{\{a;\mathfrak{R}\}} = I - [B \cdot B^\dagger \; ; \; a \cdot B = 0, \; B \in \mathcal{F}]$$

These tensors are projectors in $\mathcal{R}_{\eta\eta^*}$ and $\mathcal{F}_{\eta\eta^*}$, respectively, and are characterized by the following two properties.

B 2.1f Lemma. If $A_\theta^\eta \in \mathcal{R}$ and $B_\theta^\eta \in \mathcal{F}$ are such that $a \cdot A = 0$ and $a \cdot B = 0$, then $E_{\{a\}} \cdot A = 0$ and $E_{\{a;\mathfrak{R}\}} \cdot B = 0$.

B 2.1g Lemma. $a \cdot E_{\{a\}} = a$ and $a \cdot E_{\{a;\mathfrak{R}\}} = a$.

Proofs. If $a \cdot A = 0$, then $E_{\{a\}} \cdot A = E_{\{a\}} \cdot (A \cdot A^\dagger) \cdot A = 0$ by A4.2c . Similarly, $E_{\{a;\mathfrak{R}\}} \cdot B = 0$. Also, $a \cdot (I - E_{\{a\}}) = 0$ follows from A4.2d

so that Lemma B 2.1g follows.

Thus, $E_{\{a\}}$ is the unique projector in $\mathfrak{R}$ having the same right annhilators in $\mathfrak{R}$ as a and $E_{\{a;\mathfrak{R}\}}$ is the corresponding projector with $\mathfrak{F}$ in place of $\mathfrak{R}$.

B 2.1h Notation. If b_β, .., d_δ are elements of $\mathfrak{C}$ and r, .., t $\geqslant$ 0, I shall write

$$E_{\{(b)^\wedge..(d)^t\}} = E^{(\wedge,..,t)}_{\{b,..,d\}}$$

and $$E_{\{(b)^\wedge..(d)^t;\mathfrak{R}\}} = E^{(\wedge,..,t)}_{\{b,..,d;\mathfrak{R}\}}.$$

B 2.1i Notation. If $b'^{\beta^*} = b_\beta$, .., $d'^{\delta^*} = d_\delta$ then projectors $E^{(\wedge,..,t)}_{\{b',..,d'\}}$ and $E^{(\wedge,..,t)}_{(b',..,d';\mathfrak{R})}$ are the same as those defined in B 2.1g but with the reverse notation with respect to contracted products (e.g. $E^{(\wedge)}_{\{b'\}} \bullet (b')^\wedge = (b')^\wedge$). These projectors may also be written $E'^{(\wedge,..,t)}_{\{b,..,d\}}$ and $E'^{(\wedge,..,t)}_{\{b,..,d;\mathfrak{R}\}}$ respectively.

B 2.2 Theorem. Let b_β, .., d_δ be elements of $\mathfrak{C}$. Then for each set of integers r, .., t $\geqslant$ 0 and for each element $A^\eta_\bullet$ of $\mathfrak{R}$ [of $\mathfrak{F}$], where $\eta = \beta^\wedge..\delta^t$, the relation $\{(b)^\wedge..(d)^t\} \bullet A = 0$ holds if and only if $E^{(\wedge,..,t)}_{\{b,..,d\}} \bullet A = 0$ $\left[E^{(\wedge,..,t)}_{\{b,..,d;\mathfrak{R}\}} \bullet A = 0\right]$.

Also, the tensors $E^{(\wedge,..,t)}_{\{b,..,d\}}$ [the tensors $E^{(\wedge,..,t)}_{\{b,..,d;\mathfrak{R}\}}$] are the only projectors in $\mathfrak{R}$ [in $\mathfrak{F}$] with the above properties.

Proof. By B 2.1f, $\{(b)^\wedge..(d)^t\} \bullet A = 0$ implies $E^{(\wedge,..,t)}_{\{b,..,d\}} \bullet A = 0$. Conversely, by B 2.1g, if $E^{(\wedge,..,t)}_{\{b,..,d\}} \bullet A = 0$, then $\{(b)^\wedge..(d)^t\} \bullet A = \{(b)^\wedge..(d)^t\} \bullet E^{(\wedge,..,t)}_{\{b,..,d\}} \bullet A = 0$. The uniqueness property follows from the fact that any projector E^η_η

in $\mathfrak{R}$ is uniquely determined by its space of right annihilators in (by A 4.2) and this space is the same for $E^{(r,\cdots,t)}_{\{b,\cdots,d\}}$ as for $(b)^r \cdots (d)^t$.

The proof for $E^{(r,\cdots,t)}_{\{b,\cdots,d;\mathfrak{R}\}}$ is similar.

B 2.2a Lemma. Let $x_{(1)\alpha}, \cdots, x_{(r)\alpha}$ and $y^{\alpha}_{(1)}, \cdots, y^{\alpha}_{(r)}$ be, respectively, r linearly independent elements of $\mathcal{C}_\alpha$ and r linearly independent elements of $\mathcal{C}_{\alpha^*}$; then

$$E^{(a_1,\cdots,a_r)}_{\{x_{(1)},\cdots,x_{(r)};\mathfrak{R}\}} = E^{(a_1,\cdots,a_r)}_{\{y_{(1)},\cdots,y_{(r)};\mathfrak{R}\}} = E^{(a_1,\cdots,a_r)}_{\{x_{(1)},\cdots,x_{(r)}\}} = E^{(a_1,\cdots,a_r)}_{\{y_{(1)},\cdots,y_{(r)}\}} = P^{(a_1)}_{(\alpha)} \cdots P^{(a_r)}_{(\alpha)}$$

Proof. Let $A^{\eta}_{\theta} \in \mathcal{F}$, where $\eta = \alpha^{a_1+\cdots+a_r}$ and suppose $\{(x_{(1)})^{a_1} \cdots (x_{(r)})^{a_r}\} \cdot A = 0$. Now, let $x'_{(1)\alpha}, \cdots, x'_{(r)\alpha}$ be another set of linearly independent elements of $\mathcal{C}_\alpha$. By A 3.3m, since α is a generator of $\mathfrak{S}$, there is a linear transformation Φ which sends $x_{(i)}$ into $x'_{(i)}$ for each i. But, by A 3.3k $\Phi(A) = A$. Hence, $\{(x'_{(1)})^{a_1} \cdots (x'_{(r)})^{a_r}\} \cdot A = 0$ and this holds for any set of linearly independent elements $x'_{(1)}, \cdots, x'_{(r)}$ of $\mathcal{C}_\alpha$.

Thus it holds for any set of r elements of $\mathcal{C}_\alpha$, so that by B 1.2 $\{P^{(a_1)}_{(\alpha)} \cdots P^{(a_r)}_{(\alpha)}\} \cdot A = 0$.

Conversely, if $\{P^{(a_1)}_{(\alpha)} \cdots P^{(a_r)}_{(\alpha)}\} \cdot A = 0$, then by B 1.2, $\{(x_{(1)})^{a_1} \cdots (x_{(r)})^{a_r}\} \cdot A = 0$. But $P^{(a_1)}_{(\alpha)} \cdots P^{(a_r)}_{(\alpha)}$ is a projector in $\mathcal{F}$ so by B 2.2, $E^{(a_1,\cdots,a_r)}_{\{x_{(1)},\cdots,x_{(r)};\mathfrak{R}\}} = P^{(a_1)}_{(\alpha)} \cdots P^{(a_r)}_{(\alpha)}$. The proof for the $y_{(i)}$'s and for $E^{(a_1,\cdots,a_r)}_{\{x_{(1)},\cdots,x_{(r)}\}}$, etc., is similar.

B 2.2b Notation. Let $b_\beta, \cdots, d_\delta \in \mathcal{C}$. I shall denote by $\mathcal{V}_{\{b,\cdots,d\}}$, the geometric variety of solutions of $\{(x)^r \cdots (z)^t\} \cdot \{I - E^{(r,\cdots,t)}_{\{b,\cdots,d\}}\} = 0$ where r, .., t take all non-negative integral values. By the Hilbert basis

theorem, a finite number of such values will suffice.) This is a variety in the many-way space $\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta$. The invariant variety $\mathcal{V}_{\{b,..,d;\mathfrak{R}\}}$ is defined similarly to $\mathcal{V}_{\{b,..,d\}}$ but with $E^{(\lambda,..,t)}_{\{b,..,d;\mathfrak{R}\}}$ in place of $E^{(\lambda,..,t)}_{\{b,..,d\}}$.

B 2.2c Lemma. Let $b_\beta, .., d_\delta \in \mathcal{C}$; then $(b, .., d) \in \mathcal{V}_{\{b,..,d\}}$ and $(b, .., d) \in \mathcal{V}_{\{b,..,d;\mathfrak{R}\}}$.

Proof. By B 2.1g $\{(b)^\wedge..(d)^t\}\cdot\{I - E^{(\lambda,..,t)}_{\{b,..,d\}}\} = 0 = \{(b)^\wedge..(d)^t\}\cdot\{I - E^{(\lambda,..,t)}_{\{b,..,d;\mathfrak{R}\}}\}$ for each set of non-negative integers $(r,.., t$

B 2.2d Lemma. If $x_\beta, .., z_\delta \in \mathcal{C}$, then $(x, ..,z) \in \mathcal{V}_{\{b,..,d\}}$ $\left[(x, .., z) \in \mathcal{V}_{\{b,..,d;\mathfrak{R}\}}\right]$ if and only if, (for all sets of non-negative integers r, .., t;) x, .., z satisfy all relations of the form $\{(x)^\wedge.. (z)^t\}\cdot A = 0$ for which $A \in \mathcal{R}$ $[A \in \mathcal{P}]$ and $\{(b)^\wedge..(d)^t\}\cdot A = 0$.

Proof. If $\{(x)^\wedge..(z)^t\}\cdot A = 0$ whenever $\{(b)^\wedge..(d)^t\}\cdot A = 0$ and $A \in \mathcal{R}$, then $\{(x)^\wedge..(z)^t\}\cdot[A\cdot A^\dagger ; \{(b)^\wedge..(d)^t\}\cdot A = 0, A \in \mathcal{R}] = 0$ and conversely. But $[A\cdot A^\dagger ; \{(b)^\wedge.. (d)^t\}\cdot A = 0, A \in \mathcal{R}] = I - E^{(\lambda,..,t)}_{\{b,..,d\}}$. Thus the condition that $\{(x)^\wedge..(z)^t\}\cdot A = 0$ hold whenever $\{(b)^\wedge..(d)^t\}\cdot A = 0$ and $A \in \mathcal{R}$ is equivalent to $\{(x)^\wedge..(z)^t\}\cdot\{I - E^{(\lambda,..,t)}_{\{b,..,d\}}\} = 0$. There is also the corresponding result for $E^{(\lambda,..,t)}_{\{b,..,d;\mathfrak{R}\}}$.

B 2.2e Lemma. If $\mathcal{U}$ is a geometric [an invariant] variety in $\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta$, then $(b,..,d) \in \mathcal{U}$ if and only if $\mathcal{V}_{\{b,..,d\}} \subseteq \mathcal{U}$ $[\mathcal{V}_{\{b,..,d;\mathfrak{R}\}} \subseteq \mathcal{U}]$ $(b_\beta, .., d_\delta \in \mathcal{C})$.

Proof. Let $\mathcal{U}$ be given by $\{(x)^\wedge..(z)^t\}\cdot F = 0$, where $F \in \mathcal{R}$. Suppose $(b, .., d) \in \mathcal{U}$; then $\{(b)^\wedge..(d)^t\}\cdot F = 0$. Thus, by B 2.2d, any point

CHAPTER 6

$(x, .., z)$ of $\mathscr{V}_{\{b,..,d\}}$ also satisfies $\{(x)^r..(z)^t\}\cdot F = 0$, so that $\mathscr{V}_{\{b,..,d\}} \subseteq \mathscr{U}$. The proof for $\mathscr{V}_{\{b,..,d;\mathfrak{R}\}}$ is similar and the converse follows from B 2.2c. Thus $\mathscr{V}_{\{b,..,d\}}$ and $\mathscr{V}_{\{b,..,d;\mathfrak{R}\}}$ are, respectively, the smallest geometric and the smallest invariant variety in $\mathscr{C}_\beta \times .. \times \mathscr{C}_\delta$ which contains the point $(b, .., d)$.

B 2.2f **Lemma.** Each variety $\mathscr{V}_{\{b,..,d\}}$ [$\mathscr{V}_{\{b,..,d;\mathfrak{R}\}}$] is an irreducible geometric [invariant] variety.

Proof. Suppose $\mathscr{V}_{\{b,..,d\}} = \mathscr{U} \sim \mathscr{W}$ where $\mathscr{U}$ and $\mathscr{W}$ are geometric varieties. Then, since by B 2.2c $(b, .., d) \in \mathscr{V}_{\{b,..,d\}}$, we have either $(b,..,d) \in \mathscr{U}$ or $(b, ..,d) \in \mathscr{W}$. Hence, by B 2.2e, either $\mathscr{V}_{\{b,..,d\}} \subseteq \mathscr{U}$ or $\mathscr{V}_{\{b,..,d\}} \subseteq \mathscr{W}$. Thus the geometric variety $\mathscr{V}_{\{b,..,d\}}$ is not the union of two proper geometric subvarieties and is therefore irreducible. The proof for $\mathscr{V}_{\{b,..,d;\mathfrak{R}\}}$ is exactly similar.

In the case of geometric varieties (but not invariant varieties in general) there is the following converse to B 2.2f.

B 2.2g **Lemma** If $\mathscr{U}$ is an irreducible geometric variety in $\mathscr{C}_\beta \times .. \times \mathscr{C}_\delta$ then there exist elements $b, .., d$ of $\mathscr{C}$ such that $\mathscr{U} = \mathscr{V}_{\{b,..,d\}}$.

Proof. If $\mathscr{U}$ is an irreducible geometric variety in $\mathscr{C}_\beta \times .. \times \mathscr{C}_\delta$, it is an algebraic variety defined over $\mathfrak{R}_1$ which is irreducible over $\mathfrak{R}_1$. Hence $\mathscr{U}$ has a generic point $(b,..,d)$ over $\mathfrak{R}_1$. $((b, .., d) \in \mathscr{C}_\beta \times .. \times \mathscr{C}_\delta)$ (see Hodge and Pedoe (10) pp. 11 – 15). This means that if a polynomial over $\mathfrak{R}_1$ in the components of $b, .., d$ in the frame $\mathscr{F}$ of $\mathscr{C}$ vanishes, it also

vanishes if (b,..,d) is replaced by any other point (x,..,z) of $\mathcal{U}$ (It is for the existence of such a generic point that the infinite transcendence of $\mathfrak{R}$ over the rationals is required.) Thus, if $\{(b)^{\wedge}..(d)^{t}\}\bullet A = 0$ where $A \in \mathfrak{R}$, then $\{(x)^{\wedge}..(z)^{t}\}\bullet A = 0$ for all points (x, .., z) of $\mathcal{U}$. Hence, $\mathcal{U} \subseteq \mathcal{V}_{\{b,..,d\}}$, by B 2.2d. But $(b,..,d) \in \mathcal{U}$, so that, by B 2.2e, $\mathcal{V}_{\{b,..,d\}} \subseteq \mathcal{U}$. This proves the result.

B 2.2h Remark. The particular point (b, .., d) of B 2.2g is a generic point over $\mathfrak{R}_1$ of $\mathcal{V}_{\{b,..,d\}}$. However, it is not always true that (x,..,z) is a generic point over $\mathfrak{R}_1$ (in the conventional sense) of the variety $\mathcal{V}_{\{x,..,z\}}$; indeed, the definition of a generic point over $\mathfrak{R}_1$ is not independent of the frame $\mathcal{F}$ since a new frame $\mathcal{F}'$ may consist of elements transcendental over $\mathfrak{R}_1$ in $\mathcal{F}$. (That is, $\mathfrak{R}_1$ is not really a "ground field", but it was taken as such for the purpose of proving B 2.2g .) The reason that B 2.2g does not hold for invariant varieties is that we are not allowed to extend the field $\mathfrak{R}$ in order to define "transcendental points" over $\mathfrak{R}$.

B 2.21 Definition. The set (x, .., z) will be called a generic set of tensors of $\mathcal{V}_{\{x,..,z\}}$. Thus, any irreducible geometric variety has a generic set of tensors and, conversely, any set of tensors is generic for just one irreducible geometric variety. This may be compared with the corresponding result for generic points of algebraic varieties (see Hodge and Pedoe (10) pp.11-15).

If $a_\beta, \ldots, c_\delta$ are elements of $\mathfrak{C}$, I shall say that the element x_ζ is a <u>generic tensor of $\mathfrak{V}_{\{y\}}$ with respect to</u> a, .., c if $E^{(q,\wedge,\ldots,t)}_{\{x,a,\ldots,c\}} = E^{(q)}_{\{y\}} E^{(\wedge,\ldots,t)}_{\{a,\ldots,c\}}$ for each r, .., t,q.

B 2.2j <u>Lemma</u>. Let $a_\zeta \in \mathfrak{C}$ and $\beta, \ldots, \delta \in \mathfrak{S}$; then for each irreducible geometric variety $\mathfrak{V} = \mathfrak{V}_{\{a,\ldots,d\}}$ in $\mathfrak{C}_\beta \times \ldots \times \mathfrak{C}_\delta$, there exist points $(x_{(1)\beta}, \ldots, z_{(1)\delta})$, $(x_{(2)\beta}, \ldots, z_{(2)\delta})$, ..., $(x_{(p)\beta}, \ldots, z_{(p)\delta})$ of $\mathfrak{V}$ such that any $A^\eta_\bullet \in \mathfrak{R}$ for which

$$\{a(x_{(1)})^\wedge \ldots (z_{(1)})^t\} \bullet A = \ldots = \{a(x_{(p)})^\wedge \ldots (z_{(p)})^t\} \bullet A = 0$$

$(\eta = \beta^\wedge \ldots \delta^t \; ; \; r, \ldots, t \geqslant 0)$, also satisfies $\{E_{\{a\}} E^{(\wedge,\ldots,t)}_{\{b,\ldots,d\}}\} \bullet A = 0$.

<u>Proof</u>. The proof is similar to that of B 2.2g. Let $\lambda_1, \ldots, \lambda_m$ be the components of a, .., c in the frame $\mathfrak{F}$ of $\mathfrak{C}$. Let $\mathfrak{H}$ be the field $\mathfrak{R}_1(\lambda_1, \ldots, \lambda_m)$ and let the irreducible components of $\mathfrak{V}$ over $\mathfrak{H}$, be $\mathfrak{V}_1, \ldots, \mathfrak{V}_p$. (In fact the field $\mathfrak{H}$ and also the reducibility of $\mathfrak{V}$ over $\mathfrak{H}$ depend on the choice of $\mathfrak{F}$.) Let $(x_{(i)}, \ldots, z_{(i)})$ be a generic point over $\mathfrak{H}$ of $\mathfrak{V}_i$, i = 1, .., p ($\mathfrak{R}$ has infinite transcendence over $\mathfrak{R}_1$ and, therefore, also over $\mathfrak{H}$). Then if $\{a(x_{(i)})^\wedge \ldots (z_{(i)})^t\} \bullet A = 0$ where $A \in \mathfrak{R}$, this relation must also hold if $(x_{(i)}, \ldots, z_{(i)})$ is replaced by any other point of $\mathfrak{V}_i$. Thus, if $A \in \mathfrak{R}$ and $\{a(x_{(i)})^\wedge \ldots (z_{(i)})^t\} \bullet A = 0$ holds for each i, i = 1, .., p, then $\{a(x)^\wedge \ldots (z)^t\} \bullet A$ holds for each point (x, .., z) of $\mathfrak{V}$. Hence, by B 2.2, $\{a\, E^{(\wedge,\ldots,t)}_{\{b,\ldots,d\}}\} \bullet A = 0$ so that, by B 2.1f, $E_{\{a\}} E^{(\wedge,\ldots,t)}_{\{b,\ldots,d\}} \bullet A = 0$.

B 2.2k If x is generic with respect to a, .., c, this means that x is "independent" of a, .., c as far as homogeneous relations are concerned. If $\nu(\alpha) = 2$ and p_{α^*}, q_{α^*} are two distinct non-zero elements of $\mathcal{C}_{\alpha^*}$, then p + q is generic with respect to p and q, but pq is not. This is because pq is homogeneous but p + q is not. If x, a, .., c represent geometrical objects in the projective space $\mathfrak{P}$, then the tensor x is generic with respect to a, .., c when the object represented by x is independent of those represented by a, .., c as regards any kind of geometrical construction. Thus, p + q is generic with respect to p and q since p + q just represents some point (other than p or q) on the line $\mathfrak{P}$. However p + q is not generic with respect to p, q and p - q, since p + q is the harmonic conjugate of p - q with respect to p and q. In fact, there is the relation $(p-q)pq(p+q) \bullet A = 0$ where $A = I^{(i)} p^{(i)} I^{(i)} \bullet (N^{(i)})^2$, whereas $(p-q)pqI^{(i)} \bullet A \neq 0$ showing that $E^{(i,i,i,i)}_{\{p-q,p,q,p+q\}} \neq E^{(i,i,i)}_{\{p-q,p,q\}} E^{(i)}_{\{p+q\}}$ $(E^{(i)}_{\{p+q\}} = I^{(i)})$.

B 2.21 Lemma. Let a^ζ, b_β, .., $d_\delta \in \mathcal{C}$, where $\zeta = \beta^\wedge .. \delta^\varepsilon$ and suppose $\{(x)^\lambda .. (z)^\varepsilon\} \bullet a = 0$ for each point (x, .., z) of $\mathcal{Q}_{\{b,..,d\}}$; then $E^{(\lambda,..,\varepsilon)}_{\{b,..,d\}} \bullet a = 0$.

Proof. Writing $a'_{\zeta^*} = a^\zeta$ and $A^{\zeta^*\zeta} = I_{(\zeta^*)}$, the relation $\{(x)^\lambda .. (z)^\varepsilon\} \bullet a = 0$ becomes $\{a'(x)^\lambda .. (z)^\varepsilon\} \bullet A = 0$. Thus lemma B 2.2j applies, whence $E_{\{a'\}} E^{(\lambda,..,\varepsilon)}_{\{b,..,d\}} \bullet A = 0$; i.e. $a' E^{(\lambda,..,\varepsilon)}_{\{b,..,d\}} \bullet A = 0$; i.e. $E^{(\lambda,..,\varepsilon)}_{\{b,..,d\}} \bullet a = 0$.

B 2.2m **Lemma.** Let a^λ, b_β, .., $d_\delta \in \mathcal{C}$, where $\zeta = \beta^\wedge .. \delta^\tau$ and suppose $\{(x)^\wedge .. (z)^\tau\} \cdot a = 0$ for each point (x, .., z) of $\mathcal{V}_{\{b,..,d;\mathfrak{R}\}}$; then $E_{\{b,..,d;\mathfrak{R}\}} \cdot a = 0$.

Proof. Let Φ be any linear transformation of $\mathcal{C}$. Now, $\mathcal{V}_{\{b,..,d;\mathfrak{R}\}}$, being an invariant variety is unchanged by a linear transformation. Thus, if $\{(x)^\wedge .. (z)^\tau\} \cdot a = 0$ for every point (x, .., z) of $\mathcal{V}_{\{b,..,d;\mathfrak{R}\}}$, then $\{(x)^\wedge .. (z)^\tau\} \cdot b = 0$ where $b = \Phi(a)$. Hence, by B 1.2g, there exists a projector F of $\mathcal{F}$ such that $\{(b)^\wedge .. (c)^\tau\} \cdot F = 0$ and $F \cdot a = a$. By B 2.2, $E^{(\wedge,..,\tau)}_{\{b,..,d;\mathfrak{R}\}} \cdot F = 0$ whence $E^{(\wedge,..,\tau)}_{\{b,..,d;\mathfrak{R}\}} \cdot a = E^{(\wedge,..,\tau)}_{\{b,..,d;\mathfrak{R}\}} \cdot (F \cdot a) = 0$.

Lemmas B 2.21 and B 2.2m are essentially equivalent to the statements of the respective formulae

$$\left.\begin{aligned} E^{(\wedge,..,\tau)}_{\{b,..,d\}} &= \left[\{(x)^\wedge .. (z)^\tau\}^\dagger \{(x)^\wedge .. (z)^\tau\} \;;\; (x,..,z) \in \mathcal{V}_{\{b,..,d\}} \right] \\ E^{(\wedge,..,\tau)}_{\{b,..,d;\mathfrak{R}\}} &= \left[\{(x)^\wedge .. (z)^\tau\}^\dagger \{(x)^\wedge .. (z)^\tau\} \;;\; (x,..,z) \in \mathcal{V}_{\{b,..,d;\mathfrak{R}\}} \right] \end{aligned}\right\} \text{B 2.2n}$$

(by virtue of B 2.1g).

B 2.2o **Notation.** Let $\mathcal{U}$ be an invariant variety in $\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta$

Define

$$E^{(\wedge,..,\tau)}_{\mathcal{U}} = \left[\{(x)^\wedge .. (z)^\tau\}^\dagger \{(x)^\wedge .. (z)^\tau\} \;;\; (x, .., z) \in \mathcal{U} \right]$$

where r, .., t are non-negative integers.

By B 2.2n, we have

$$E^{(\wedge,..,\tau)}_{\mathcal{V}_{\{b,..,d\}}} = E^{(\wedge,..,\tau)}_{\{b,..,d\}} \text{ and } E^{(\wedge,..,\tau)}_{\mathcal{V}_{\{b,..,d;\mathfrak{R}\}}} = E^{(\wedge,..,\tau)}_{\{b,..,d;\mathfrak{R}\}} \qquad \text{B 2.2p}$$

B 2.2q **Lemma.** Let $\mathcal{V}_i$ ($i \in \Omega$) be some set of invariant varieties in $\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta$ and let $\mathcal{U}$ be their union; then $E^{(\wedge,..,\tau)}_{\mathcal{U}} = [E^{(\wedge,..,\tau)}_{\mathcal{V}_i} \;;\; i \in \Omega]$

for each set of non-negative integers (r, .., t).

Proof. The set of right annihilators in $\mathcal{C}$ of $E_{\mathcal{U}}^{(r,..,t)}$ is the intersection of the sets of right annihilators in $\mathcal{C}$ of the tensors $(x)^r .. (z)^t$ for which $(x, .., z) \in \mathcal{V}_i$. Thus, it is the set of elements of $\mathcal{C}$ which are right annihilators of all those tensors $(x)^r .. (z)^t$ for which $(x, .., z) \in \mathcal{V}_i$, for all $i \in \Omega$. It is therefore the intersection of the sets of right annihilators of each $E_{\mathcal{V}_i}^{(r,..,t)}$ where $i \in \Omega$. Hence $E_{\mathcal{U}}^{(r,..,t)} = [E_{\mathcal{V}_i}^{(r,..,t)} ; i \in \Omega]$.

B 2.2r Lemma. Let $\mathcal{U}$ be an invariant variety in $\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta$; then $E_{\mathcal{U}}^{(r,..,t)}$ is an element of $\mathcal{F}$, for each r, .., $t \geqslant 0$. Also, each tensor $E_{\mathcal{U}}^{(r,..,t)}$ is an element of $\mathcal{R}$ if $\mathcal{U}$ is a geometric variety.

Proof. By B 2.2e, if $(x, .., z) \in \mathcal{U}$, then $\mathcal{V}_{\{x,..,z;\mathfrak{R}\}} \subseteq \mathcal{U}$. Also, $(x, .., z) \in \mathcal{V}_{\{x,..,z;\mathfrak{R}\}}$ by B 2.2c. Thus $\mathcal{U}$ is the union of the $\mathcal{V}_{\{x,..,z;\mathfrak{R}\}}$ for which $(x, .., z) \in \mathcal{U}$. Similarly, if $\mathcal{U}$ is geometric, then, by B 2.2e and B 2.2c, it follows that $\mathcal{U}$ is the union of the $\mathcal{V}_{\{x,..,z\}}$ for which $(x, .., z) \in \mathcal{U}$. Hence, by B 2.2q and B 2.2p we have,

$$E_{\mathcal{U}}^{(r,..,t)} = [E_{\{x,..,z;\mathfrak{R}\}}^{(r,..,t)} ; \ (x, .., z) \in \mathcal{U}] \qquad \text{B 2.2s}$$

and, for a geometric variety,

$$E_{\mathcal{U}}^{(r,..,t)} = [E_{\{x,..,z\}}^{(r,..,t)} ; \ (x, .., z) \in \mathcal{U}] \qquad \text{B 2.2t}$$

The result follows from this.

B 2.2u Lemma. Let $\mathcal{U}$ and $\mathcal{V}$ be invariant varieties in $\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta$; then $\mathcal{U} \subseteq \mathcal{V}$ if and only if $E_{\mathcal{U}}^{(r,..,t)} \cdot E_{\mathcal{V}}^{(r,..,t)} = E_{\mathcal{U}}^{(r,..,t)}$ for all r, .., $t \geqslant 0$.

Proof. By B 2.2q, we have $E_{\mathcal{V}}^{(r,\cdots,t)} = \left[E_{\mathcal{U}}^{(r,\cdots,t)}, E_{\mathcal{U}}^{(r,\cdots,t)} \right]$ if $\mathcal{V} = \mathcal{U}\mathcal{U}$ (see A 4.3a)

B 2.3 **Theorem.** Let $\mathcal{U}$ be a geometric [invariant] variety in $\mathcal{C}_\beta \times \cdots \times \mathcal{C}_\delta$. Let $\eta = \beta \wedge \cdots \delta^t$ where r, .., t $\geqslant 0$; then if $a_\eta \in \mathcal{C}$, $\{(x)^r \cdots (z)^t\} \cdot a = 0$ for all points $(x_\beta, .., z_\delta)$ of $\mathcal{U}$ if and only if $E_{\mathcal{U}}^{(r,\cdots,t)} \cdot a = 0$.

The projector $E_{\mathcal{U}}^{(r,\cdots,t)}$, for each r, .., t, is the only projector with the above property.

Proof. By the definition B2.2o of $E_{\mathcal{U}}^{(r,\cdots,t)}$, it follows that the right annihilators of $E_{\mathcal{U}}^{(r,\cdots,t)}$ are just those common to all $(x)^r \cdots (z)^t$ for which $(x, .., z) \in \mathcal{U}$. Thus $E_{\mathcal{U}}^{(r,\cdots,t)} \cdot a = 0$ if and only if $\{(x)^r \cdots (z)^t\} \cdot a = 0$ for all (x, .., z) of $\mathcal{U}$.

$E_{\mathcal{U}}^{(r,\cdots,t)}$ is unique by A 4.2.

B 2.3a **Corollary.** Let $(x_\beta, .., z_\delta)$ be a point of the invariant variety $\mathcal{U}$ in the space $\mathcal{C}_\beta \times \cdots \times \mathcal{C}_\delta$ then $\{(x)^r \cdots (z)^t\} \cdot E_{\mathcal{U}}^{(r,\cdots,t)} = (x)^r \cdots (z)^t$ for each r, .., t $\geqslant 0$.

Proof. Since $E_{\mathcal{U}}^{(r,\cdots,t)}$ is a projector, $E_{\mathcal{U}}^{(r,\cdots,t)} \cdot \{I - E_{\mathcal{U}}^{(r,\cdots,t)}\} = 0$. Hence, by B 2.3, if $(x, .., z) \in \mathcal{U}$, then $\{(x)^r \cdots (z)^t\} \cdot \{I - E_{\mathcal{U}}^{(r,\cdots,t)}\} = 0$ proving the result. This is also a simple consequence of the definition of $E_{\mathcal{U}}^{(r,\cdots,t)}$.

B 2.3b **Corollary.** Let $\{(x)^r \cdots (z)^t\} \cdot a$ be a function of the variable tensors $x_\beta, .., z_\delta$ whose range of variation is that $(x, .., z) \in \mathcal{U}$, $\mathcal{U}$ being some invariant variety in $\mathcal{C}_\beta \times \cdots \times \mathcal{C}_\delta$. This function then uniquely determines the tensor $b = E_{\mathcal{U}}^{(r,\cdots,t)} \cdot a$.

Also, $\{(x)^r \cdots (z)^t\} \cdot a = \{(x)^r \cdots (z)^t\} \cdot b$ for all $(x, .., z) \in \mathcal{U}$.

Proof. Suppose a' is any tensor for which $\{(x)^{\lambda}..(z)^{t}\}\bullet a' = \{(x)^{\lambda}..(z)^{t}\}\bullet a$ for all $(x, .., z) \in \mathcal{U}$; then $\{(x)^{\lambda}..(z)^{t}\}\bullet(a' - a) = 0$ for all $(x, .., z) \in \mathcal{U}$ so that, by theorem B 2.3, $E_{\mathcal{U}}^{(\lambda,..,t)}\bullet(a' - a) = 0$.

Thus $b = E_{\mathcal{U}}^{(\lambda,..,t)}\bullet a = E_{\mathcal{U}}^{(\lambda,..,t)}\bullet a'$.

That $\{(x)^{\lambda}.. (z)^{t}\}\bullet a = \{(x)^{\lambda}..(z)^{t}\}\bullet b$ follows from B 2.3 a.

B 2.3c Theorem B 2.3 and corollary B 2.3b are the respective generalizations of theorem B 1.2 and corollary B 1.2a to functions of tensors which are subject to an invariant relation. (Corollary B 2.3b gives a canonical form for such functions.) Thus, if a class of geometrical objects is represented by, say, tensors belonging to a certain geometric variety, then by B 2.3b, any homogeneous polynomial in the components of such a tensor corresponds uniquely to a tensor like the tensor b of B 2.3b so that properties of such polynomials can be studied by studying, instead, the tensor b. A particular example of this is polynomials in Grassmann coordinates. These will be considered in more detail later.

Theorem B 1.2 is a special case of B 2.3 since $E_{\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta}^{(\lambda,..,t)} = P_{(\beta)}^{(\lambda)}.. P_{(\delta)}^{(t)}$ (see B 2.3d).

B 2.3d **Lemma.** Let $\mathcal{U}_1, .., \mathcal{U}_m$ be invariant varieties. Then, if $p, .., r, ..., t, .., v \geqslant 0$,

$$E_{\mathcal{U}_1}^{(p,..,\lambda)} ... E_{\mathcal{U}_m}^{(\iota,..,\nu)} = E_{\mathcal{U}_1 \times .. \times \mathcal{U}_m}^{(p,..,\lambda,\,.....,\,t,..,\nu)}.$$

Proof. It is sufficient to consider only the case $m = 2$ since the general case will follow from this.

Let $\mathcal{U}_1 \subseteq \mathcal{C}_\alpha \times .. \times \mathcal{C}_\delta$, $\mathcal{U}_2 \subseteq \mathcal{C}_\zeta \times .. \times \mathcal{C}_\theta$ and let $a_\xi^{\alpha .. \delta \wedge \zeta^\tau .. \theta^\nu} \in \mathcal{C}$.

Then if $x_\alpha, .., z_\delta, x'_\zeta, .., z'_\theta \in \mathcal{C}$, we have, by B 2.3, $E_{\mathcal{U}_1}^{(\rho,..,\lambda)} E_{\mathcal{U}_2}^{(\tau,..,\nu)} \cdot a = 0$ if and only if $\{(x)^\rho .. (z)^\lambda\} E_{\mathcal{U}_2}^{(\tau,..,\nu)} \cdot a = 0$ for each $(x, .., z) \in \mathcal{U}_1$. That is, if and only if $\{(x)^\rho .. (z)^\lambda\}\{(x')^\tau .. (z')^\nu\} \cdot a = 0$ for each $(x, .., z) \in \mathcal{U}_1$, $(x', .., z') \mathcal{U}_2$, i.e. for each $(x, .., z, x', .., z') \in \mathcal{U}_1 \times \mathcal{U}_2$. But $E_{\mathcal{U}_1}^{(\rho,..,\lambda)} E_{\mathcal{U}_2}^{(\tau,..,\nu)}$ is a projector (by A 4.1a), whence, by B 2.3, $E_{\mathcal{U}_1 \times \mathcal{U}_2}^{(\rho,..,\lambda,\tau,..,\nu)} = E_{\mathcal{U}_1}^{(\rho,..,\lambda)} E_{\mathcal{U}_2}^{(\tau,..,\nu)}$ as required.

B 2.3e **Lemma.** Let $\mathcal{U}$ be an irreducible invariant [geometric] variety in $\mathcal{C}_\alpha \times .. \times \mathcal{C}_\delta$. Let $a^{\alpha \wedge .. \delta^\tau}$, $b^{\alpha \wedge .. \delta^\tau}$ be non-zero elements of $\mathcal{C}_{\alpha^* \wedge .. \delta^{*\tau}}$ [with rational components in some frame]. Let $\mathcal{H}$ and $\mathcal{K}$ be the primals in $\mathcal{C}_\alpha \times .. \times \mathcal{C}_\delta$ whose points $(x_\alpha, .., z_\delta)$ are given by the respective equations

$$(x)^\lambda .. (z)^\tau \cdot a = 0 \text{ and } (x)^\lambda .. (z)^\tau \cdot b = 0.$$

Suppose that $\mathcal{H}$ and $\mathcal{K}$ have the same multiplicative intersection with $\mathcal{U}$. Then there is a non-zero scalar λ such that

$$E_{\mathcal{U}}^{(\lambda,..,\tau)} \cdot a = \lambda \, E_{\mathcal{U}}^{(\lambda,..,\tau)} \cdot b.$$

(If $\mathcal{H} \frown \mathcal{U}$ and $\mathcal{K} \frown \mathcal{U}$ are irreducible [over the rationals] then it is sufficient to suppose that they are the same point sets.)

Proof. The primals whose equations are

$$(x)^\lambda .. (z)^\tau \cdot a - \lambda (x)^\lambda .. (z)^\tau \cdot b = 0$$

for each non-zero $\lambda \in \mathfrak{R}$ constitute, together with $\mathcal{H}$ and $\mathcal{K}$, the pencil determined by $\mathcal{H}$ and $\mathcal{K}$. Now since $\mathcal{U}$ is irreducible over $\mathfrak{R}$ [over $\mathfrak{R}_1$], one (or all) of these must contain $\mathcal{U}$. Thus, for some non-zero λ,

$$(x)^{\wedge} .. (z)^{\dagger} \cdot a = \lambda (x)^{\wedge} .. (z)^{\dagger} \cdot b$$

for all $(x, .., z) \in \mathcal{U}$, so that the result follows from B 2.3.

B 2.3f A special case of B 2.3e, for example, would be $\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{U}_3$ where $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3$ are invariant varieties of elements x_β, y_γ, z^δ in $\mathcal{C}_\beta, \mathcal{C}_\gamma, \mathcal{C}_{\delta*}$, respectively. Then if $a = a_{\delta\epsilon}^{\beta\wedge\gamma\wedge}$, $b = b_{\delta\epsilon}^{\beta\wedge\gamma\wedge}$, the conclusion of the lemma would take the form

$$\mathbf{E}^{(\wedge)}_{\mathcal{U}_1} \mathbf{E}^{(a)}_{\mathcal{U}_2} \cdot a \cdot \mathbf{E}^{(z)}_{\mathcal{U}_3} = \lambda \mathbf{E}^{(\wedge)}_{\mathcal{U}_1} \mathbf{E}^{(a)}_{\mathcal{U}_2} \cdot b \cdot \mathbf{E}^{(z)}_{\mathcal{U}_3}$$

(see B 2.3d). The proof of B 2.3e would also be valid in terms of the many-way space $|\mathcal{C}_\beta| \times .. \times |\mathcal{C}_\delta|$ instead of the affine space $\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta$.

B 2.4 Basis relations for invariant varieties.

The results of this subsection will be concerned with varieties in the space $\mathcal{C}_\beta$. However, these results will apply equally well (with suitable modifications) to varieties in $\mathcal{C}_\beta \times .. \times \mathcal{C}_\delta$, the simpler case being taken purely for reasons of notational convenience.

B 2.4a Definition. Let $x_\beta \in \mathcal{C}$ and let $\mathcal{U}$ be an invariant variety in $\mathcal{C}_\beta$. The relations $(x)^{\wedge_i} \cdot b_{(i)} = 0$ $(i = 1, .., m)$ where $b_{(i)} \in \mathcal{C}$, will be said to form a basis for $\mathcal{U}$ if for each $a_\gamma^{\beta\wedge} \in \mathcal{C}$ such that $(x)^{\wedge} \cdot a = 0$ for all $x \in \mathcal{U}$, there exist elements $w_{(i)}$ of $\mathcal{C}$ satisfying

$$P^{(\lambda)}_{(\beta)} \bullet a = P^{(\lambda)}_{(\beta)} \bullet \sum_{\substack{i=1 \\ \lambda_i \leqslant \lambda}}^{m} \left\{ \left(b_{(i)} \, I^{(\lambda-\lambda_i)}_{(\beta)} \right) \bullet w_{(i)} \right\}.$$

This is equivalent to the existence of elements $u_{(ij)}$, $v_{(i,j)} \in \mathcal{C}$ $(i = 1, .., m ; j = 1, .., l)$ such that $(x)^{\lambda} \bullet a = \sum_{\substack{i,j \\ \lambda_i \leqslant \lambda}} \{(X)^{\lambda_i} \bullet b_{(i)} \bullet u_{(ij)}\} \{(X)^{\lambda-\lambda_i} \bullet w_{(i,j)}\}$. for all $x_\beta \in \mathcal{C}_\beta$. Hence the polynomials given by $(x)^{\lambda_i} \bullet b_{(i)}$ constitute a basis for the ideal of polynomial vanishing on $\mathcal{U}$.

B 2.4b Lemma. If $\mathcal{U}$ is an invariant variety in $\mathcal{C}_\beta$, and m is sufficiently large, the relations $(x)^i \bullet \{I - E^{(i)}_{\mathcal{U}}\} = 0$ $(i = 1, 2, \ldots m)$ form a basis for $\mathcal{U}$.

Proof. The condition for a basis is trivially satisfied since, if $(x)^i \bullet a = 0$ for all $x \in \mathcal{U}$, then $E^{(i)}_{\mathcal{U}} \bullet a = 0$, so that $P^{(\lambda)}_{(\beta)} \bullet a = P^{(\lambda)}_{(\beta)} \bullet \{I - E^{(i)}_{\mathcal{U}}\} \bullet a$. Also, by the Hilbert basis theorem, a finite number of the relations $(x)^i \bullet \{I - E^{(i)}_{\mathcal{U}}\} = 0$ is sufficient.

B 2.4c Lemma. Suppose the relations $(x)^{\lambda_i} \bullet b_{(i)} = 0$, $(i = 1, .., m)$ form a basis for $\mathcal{U}$. Let $e^{(\lambda)}_{(i)} = I^{(\lambda)}_{(\beta)} - \{b_{(i)} \bullet b^{\dagger}_{(i)}\} I^{(\lambda-\lambda_i)}_{(\beta)}$ for each $r \geqslant 0$ for which $r \geqslant r_i$ and let $e^{(\lambda)}_{(i)} = I^{(\lambda)}_{(\beta)}$ otherwise. Then $E^{(\lambda)}_{\mathcal{U}} = (P^{(\lambda)}_{(\beta)}, e^{(\lambda)}_{(1)}, \ldots, e^{(\lambda)}_{(m)})$ for each $r \geqslant 0$.

Conversely, if $e^{(\lambda)}_{(i)}$ and $b_{(i)}$ satisfy the above formulae for $i = 1, .., m$, then the relations $(x)^{\lambda_i} \bullet b_{(i)} = 0$ $(i = 1, .., m)$ form a basis for $\mathcal{U}$.

Proof. By B 2.3, each $E^{(\lambda)}_{\mathcal{U}}$ is uniquely determined by the fact that it is a projector and that $E^{(\lambda)}_{\mathcal{U}} \bullet a = 0$ if and only if $(x)^{\lambda} \bullet a = 0$ for all $x \in \mathcal{U}$. Put $F^{(\lambda)} = I - (P^{(\lambda)}, e^{(\lambda)}_{(1)}, \ldots, e^{(\lambda)}_{(m)})$. Then the first part of B 2.4c will

follow if it can be shown that $\mathbf{F}^{(\lambda)}\bullet a = a$ if and only if, for all $x \in \mathcal{U}$, $(x)^{\lambda}\bullet a = 0$. Hence, both parts of B 2.4c will follow if it can be shown that $\mathbf{F}^{(\lambda)}\bullet a = a$ if and only if $P^{(\lambda)}\bullet a$ is of the form $P^{(\lambda)}\bullet \sum_{\lambda_i \leq \lambda} \{(b_{(i)} I^{(\lambda-\lambda_i)})\bullet w_{(i)}\}$.

Suppose, first, that $P^{(\lambda)}\bullet a = P^{(\lambda)}\bullet \{\Sigma (b_{(i)} I^{(\lambda-\lambda_i)})\bullet w_{(i)}\}$.

Now, $I - \mathbf{F}^{(\lambda)} = (P^{(\lambda)}, e^{(\lambda)}_{(1)}, .., e^{(\lambda)}_{(m)})$ so that, by A 4.2d, $(I - \mathbf{F}^{(\lambda)}) = (I - F^{(\lambda)})\bullet P^{(\lambda)}$. Also, by A 4.2d, $(I - \mathbf{F}^{(\lambda)}) = (I - \mathbf{F}^{(\lambda)})\bullet e^{(\lambda)}_{(i)}$. But, if $r \geqslant r_i$, $e^{(\lambda)}_{(i)}\bullet (b_{(i)} I^{(\lambda-\lambda_i)}) = 0$. Hence, $(I - \mathbf{F}^{(\lambda)})\bullet \sum_{\lambda_i \leq \lambda} \{(b_{(i)} I^{(\lambda-\lambda_i)})\bullet w_{(i)}\} = 0$, so that $(I - \mathbf{F}^{(\lambda)})\bullet a = (I - \mathbf{F}^{(\lambda)})\bullet P^{(\lambda)}\bullet a = (I - \mathbf{F}^{(\lambda)})\bullet P^{(\lambda)}\bullet \sum \{(b_{(i)} I^{(\lambda-\lambda_i)})\bullet w_{(i)}\}$
$= (I - \mathbf{F}^{(\lambda)})\bullet \sum \{(b_{(i)} I^{(\lambda-\lambda_i)})\bullet w_{(i)} = 0$.

Conversely, suppose that $a = \mathbf{F}^{(\lambda)} a$. Now, by A 4.2b,

$\mathbf{F}^{(\lambda)} = [I - P^{(\lambda)}, I - e^{(\lambda)}_{(1)}, .., I - e^{(\lambda)}_{(m)}]$
$= \{I - P^{(\lambda)} + \sum_{\lambda_i \leq \lambda} (b_{(i)}\bullet b^{\dagger}_{(i)}) I^{(\lambda-\lambda_i)}\}\bullet\{I - P^{(\lambda)} + \sum_{\lambda_i \leq \lambda} (b_{(i)}\bullet b^{\dagger}_{(i)}) I^{(\lambda-\lambda_i)}\}^{\dagger}$

Thus, putting $v_{(i)} = (b^{\dagger}_{(i)} I^{(\lambda-\lambda_i)})\bullet \{I - P^{(\lambda)} + \sum_{\lambda_i \leq \lambda} (b_{(i)}\bullet b^{\dagger}_{(i)}) I^{(\lambda-\lambda_i)}\}^{\dagger}$ we get $\mathbf{F}^{(\lambda)} = (I - P^{(\lambda)})\bullet \{...\}^{\dagger} + \sum \{(b_{(i)} I^{(\lambda-\lambda_i)})\bullet v_{(i)}\}$ so that $P^{(\lambda)}\bullet a = P^{(\lambda)}\bullet \mathbf{F}^{(\lambda)}\bullet a = P^{(\lambda)}\bullet \sum \{(b_{(i)} I^{(\lambda-\lambda_i)})\bullet (v_{(i)}\bullet a)$.

B 2.4d <u>Lemma</u>. Let $\mathcal{U}$ be an invariant variety in $\mathcal{C}_{\beta}$ and let $A^{\beta_1 \cdots \beta_\lambda \beta^*_1 \cdots \beta^*_{\lambda-s}}_{\beta_{\lambda+1} \cdots \beta_\lambda} = E^{(\lambda)\beta_1 \cdots \beta_\lambda}_{\mathcal{U}\ \beta_1 \cdots \beta_\lambda}$; $r \geqslant s \geqslant 0$. Then $E^{(\lambda)}_{\mathcal{U}} = A^{\dagger}\bullet A$.

<u>Proof</u>. By B 2.3, $E^{(\lambda)}_{\mathcal{U}}$ is the unique projector whose right annihilators in $\mathcal{C}$ are those common to all tensors $(x)^{\lambda}$ for which $x \in \mathcal{U}$. But, by the definition of A, the tensor a is a right annihilator of A if and only if $(I^{(\lambda-\lambda)}_{(\beta)} a)$ is a right annihilator of $E^{(\lambda)}_{\mathcal{U}}$; that is, if and only if $(x)^{\lambda}\bullet (I^{(\lambda-\lambda)}_{(\beta)} a) = 0$ for all x of $\mathcal{U}$. But $(x)^{\lambda}\bullet (I^{(\lambda-\lambda)}_{(\beta)} a) = (x)^{\lambda-\lambda}\{(x)^{\lambda}\bullet a\}$ and

this is zero if and only if $(x)^a \cdot a = 0$. Thus, the right annihilators of A are those tensors a of $\mathscr{C}$ for which $(x)^a \cdot a = 0$ for all x of $\mathscr{U}$, i.e. they are the right annihilators of $E_{\mathscr{U}}^{(a)}$. Hence (by A 4.1c) $E_{\mathscr{U}}^{(a)} = A^\dagger \cdot A$.

B 2.4e Lemma. If $r \geqslant s \geqslant 0$, $E_{\mathscr{U}}^{(s)} \cdot (E_{\mathscr{U}}^{(a)} I_{(\beta)}^{(r-a)}) = E_{\mathscr{U}}^{(r)} = E_{\mathscr{U}}^{(r)} \cdot P_{(\beta)}^{(r)}$.

Proof. Defining A as in B 2.4d we have $E_{\mathscr{U}}^{(a)} = A^\dagger \cdot A$. The equation $E_{\mathscr{U}}^{(r)} \cdot \{(A^\dagger \cdot A) I^{(r-a)}\} = E_{\mathscr{U}}^{(r)}$ is then, simply the relation $A \cdot A^\dagger \cdot A = A$ in a different notation. Also, $E_{\mathscr{U}}^{(r)} = E_{\mathscr{U}}^{(r)} \cdot P_{(\beta)}^{(r)}$ follows from B 2.4c and B 2.3a.

B 2.4f Lemma. For each invariant variety $\mathscr{U}$ in $\mathscr{C}_\beta$ there exists an integer $m \geqslant 0$ such that $E_{\mathscr{U}}^{(m+s)} = (P_{(\beta)}^{(m+s)}, E_{\mathscr{U}}^{(m)} I_{(\beta)}^{(s)})$ for each $s \geqslant 0$.

Proof. Applying B 2.4b to B 2.4c by putting $b_{(i)} = I - E_{\mathscr{U}}^{(i)}$ $(i = 1, .., m)$, we get $e_{(i)}^{(m+s)} = I_{(\beta)}^{(m+s)} - \{I_{(\beta)}^{(i)} - E_{\mathscr{U}}^{(i)}\} I_{(\beta)}^{(m+s-i)} = E_{\mathscr{U}}^{(i)} I_{(\beta)}^{(m+s-i)}$, so that $E_{\mathscr{U}}^{(m+s)} = (P_{(\beta)}^{(m+s)}, E_{\mathscr{U}}^{(1)} I_{(\beta)}^{(m+s-1)}, \ldots, E_{\mathscr{U}}^{(m)} I_{(\beta)}^{(s)})$. But, by B 2.4e, $(E_{\mathscr{U}}^{(m)} I_{(\beta)}^{(s)}) \cdot (E_{\mathscr{U}}^{(m-r)} I_{(\beta)}^{(s+r)}) = E_{\mathscr{U}}^{(m)} I_{(\beta)}^{(s)}$ if $r \geqslant 0$. Hence the result follows (see A 4.3a).

B 2.4g Lemmas B 2.4d and B 2.4f show that the tensors $E_{\mathscr{U}}^{(r)}$ $(r = 1, 2, \ldots)$ can all be calculated if one of them $E_{\mathscr{U}}^{(m)}$, where m is sufficiently large, is known. By B 2.3, we see that the invariant variety $\mathscr{U}$ is determined by the sequence of projectors $E_{\mathscr{U}}^{(1)}$, $E_{\mathscr{U}}^{(2)}$, ... and therefore, by $E_{\mathscr{U}}^{(m)}$. In fact $\mathscr{U}$ is the set of elements x_β satisfying $(x)^m \cdot \{I - E_{\mathscr{U}}^{(m)}\} = 0$.

B 2.4h Polynomial ideals. Most of the results B 2.2o, .., B 2.4f will also apply to any algebraic varieties given by homogeneous equations. (However, the projectors $E_{\mathscr{U}}^{(r_1, \ldots, t)}$ which then arise are generally not elements of $\mathscr{T}$.)

Furthermore, it is possible to represent any homogeneous polynomial ideal by a set of projectors which satisfy B 2.4e.

Let $e^{(0)}(=1)$, $e^{(1)\beta^{\wedge}}_{\beta^{\wedge}}$, $e^{(2)\beta^{\wedge}_{1}\wedge}_{\beta^{\wedge}_{1}\wedge}$, ... be a sequence of projectors satisfying $e^{(\wedge)}\bullet(e^{(a)}\, I^{(\wedge-a)}_{(\beta)}) = e^{(\wedge)} = e^{(\wedge)}\bullet P^{(\wedge)}_{(\beta)}$ for all $r \geqslant s \geqslant 0$. Now, any expression $(x)^{\wedge}\bullet a$ where x_{β}, $a^{\beta^{\wedge}} \in \mathcal{C}$, is a homogeneous polynomial over $\mathfrak{R}$ in the $\nu(\beta)$ components of x. These polynomials correspond uniquely to the tensors $P^{(\wedge)}_{(\beta)}\bullet a$ (by B 1.2a). Let $\mathfrak{Z}$ be the set of such polynomials for which $e^{(\wedge)}\bullet a = 0$. (This is a property of the polynomial and not just of the tensor a, because $e^{(\wedge)}\bullet a = e^{(\wedge)}\bullet\{P^{(\wedge)}_{(\beta)}\bullet a\}$.)

I now show that $\mathfrak{Z}$ is a homogeneous polynomial ideal. Suppose $(x)^{\wedge}\bullet a \in \mathfrak{Z}$ and $(x)^{\wedge}\bullet b \in \mathfrak{Z}$; then clearly, $e^{(\wedge)}\bullet(a + b) = e^{(\wedge)}\bullet a + e^{(\wedge)}\bullet b = 0$ so that $(x)^{\wedge}\bullet(a + b) \in \mathfrak{Z}$. Suppose $(x)^{\wedge}\bullet a \in \mathfrak{Z}$ and $b^{\beta^{\wedge}} \in \mathcal{C}$; then $e^{(\wedge+a)}\bullet ab = e^{(\wedge+a)}\bullet(e^{(\wedge)}\, I^{(a)})\bullet ab = e^{(\wedge+a)}\bullet(e^{(\wedge)}\bullet a)b = 0$ so that $(x)^{(\wedge+a)}\bullet ab = \{(x)^{\wedge}\bullet a\}\{(x)^{a}\bullet b\} \in \mathfrak{Z}$. Thus $\mathfrak{Z}$ is a homogeneous ideal.

The results of B 2.4b, B 2.4c and B 2.4f can all be adapted so as to apply to the projectors $e^{(i)}$ corresponding to a homogeneous ideal. (B 2.4f is, there ore, a consequence of B 2.4e.) Also, it can be shown that if $\mathfrak{Z}$ and $\mathfrak{Y}$ are two ideals whose corresponding sets of projectors are $e^{(0)}$, $e^{(1)}$, $e^{(2)}$, ... and $f^{(0)}$, $f^{(1)}$, $f^{(2)}$, ..., respectively, then the projectors corresponding to the intersection and join of $\mathfrak{Z}$ and $\mathfrak{Y}$ are given by $[e^{(i)}, f^{(i)}]$ and by $(e^{(i)}, f^{(i)})$, respectively.

The following result gives the <u>postulation formula</u> (or Hilbert

function) for $\mathcal{U}$ or for $|\mathcal{U}|$ (see Hodge and Pedoe (10) p.371).

B 2.4i Lemma. Let $\mathcal{U}$ be an invariant variety in $\mathcal{C}_\beta$. Then the number of linearly independent conditions an element $a^{\beta^\wedge}$ of $\mathcal{C}_{\beta^{*\wedge}}$ has to satisfy in order that $(x)^\wedge \bullet a$ vanishes for all x_β of $\mathcal{U}$, is the integer $\langle E^{(\wedge)}_{\mathcal{U}} \rangle$.

Proof. By B 2.3a, $(x)^\wedge \bullet a = 0$ for all x of $\mathcal{U}$ if and only if $E^{(\wedge)}_{\mathcal{U}} \bullet a = 0$. The number of linearly independent tensors $a^{\beta^\wedge}$ satisfying $E^{(\wedge)}_{\mathcal{U}} \bullet a = 0$ is $\nu(\beta^\wedge) - \langle E^{(\wedge)}_{\mathcal{U}} \rangle$ by A 4.2a. But the number of linearly independent elements of $\mathcal{C}_{\beta^{*\wedge}}$ is $\nu(\beta^\wedge)$, whence the result follows.

B 2.5 An example of a set of geometric and invariant varieties.

Let $\beta = \alpha^2$ and $\nu(\alpha) = 2$; then $\nu(\beta) = 4$ so that $\mathcal{C}_\beta$ is a vector space of dimension 4 and $|\mathcal{C}_\beta|$ is a three-dimensional projective space.

Now, there are two obvious geometric varieties in $\mathcal{C}_\beta$, namely the space $\mathcal{L}$ of symmetric and the space $\mathcal{S}$ of skew-symmetric elements $a_{\alpha_1\alpha_2} = a_\beta$ of $\mathcal{C}_\beta$. The tensor a is symmetric if $a \bullet P^{(2)}_{(\alpha)} = a$ and skew symmetric of $a \bullet N^{(2)}_{(\alpha)} = a$. These are linear relations, so that $\mathcal{L}$ and $\mathcal{S}$ are linear spaces. The dimension of $\mathcal{L}$ is $\langle P^{(2)} \rangle = \frac{2.3}{1.2} = 3$ and that of $\mathcal{S}$ is $\langle N^{(2)} \rangle = \frac{2.1}{1.2} = 1$ (by A 1.5u) ; that is, $\lambda = |\mathcal{L}|$ is a plane and $S = |\mathcal{S}|$ a point. We have $E^{(\wedge)}_{\mathcal{L}} = \{P^{(2)}\}^\wedge \bullet P^{(\wedge)}_{(\beta)}$ and $E^{(\wedge)}_{\mathcal{S}} = \{N^{(2)}\}^\wedge \bullet P^{(\wedge)}_{(\beta)}$.

There is also the geometric variety $\mathcal{Q}$ of tensors $a_{\alpha_1\alpha_2}$ which factorize into two vectors ; i.e. $a_{\alpha_1\alpha_2} = u_{\alpha_1} v_{\alpha_2}$. This is a set of tensors $a_{\alpha_1\alpha_2}$ satisfying $a_{\alpha_1\alpha_2} a_{\alpha_3\alpha_4} = a_{\alpha_1\alpha_4} a_{\alpha_3\alpha_2}$.

Thus $\Sigma = |\mathcal{Q}|$ is a quadric surface in $|\mathcal{C}_\beta|$. It is non-singular and does not contain the point S as can easily be verified. Also, Σ and λ are pole and polar with respect to Σ. It can be shown that $E^{(\lambda)}_{\mathcal{Q}} = D^{(t;\lambda)} \cdot (P^{(\lambda)})^2 \cdot D^{(t;\lambda)*}$. The geometric variety which is the intersection of $\mathcal{Q}$ with $\mathcal{L}$ is the set $\mathcal{K}$ of tensors $a_{\alpha_1 \alpha_2}$ of the form $u_{\alpha_1} u_{\alpha_2}$ (that is, the square of a vector). The variety $k = |\mathcal{K}|$ is a proper conic (the "Veronesian" of binary quadratics $a \cdot (x)^2$) and $E^{(\lambda)}_{\mathcal{K}} = P^{(2\lambda)}$.

The invariant varieties in $|\mathcal{C}_\beta|$ are those algebraic varieties in which are left invariant by every linear transformation of $\mathcal{C}$. Now any such linear transformation Φ induces a linear transformation on $|\mathcal{C}_\beta|$. It leaves the point S invariant and induces a projectivity on k. It thus leaves the cone Γ projecting k from S, invariant so that Γ is an invariant variety of $|\mathcal{C}_\beta|$. In fact, Γ is a geometric variety. Now, the quadrics Γ, Σ and the plane λ counted twice are three quadrics of a pencil (Δ). Thus, every quadric of (Δ) is left invariant by Φ and is therefore an invariant variety of $|\mathcal{C}_\beta|$. Also, it turns out that every invariant variety of $|\mathcal{C}_\beta|$ has as irreducible components, quadrics of (Δ), the plane λ, the point S, the conic k, or the space $|\mathcal{C}_\beta|$ itself.

The irreducible geometric varieties of $|\mathcal{C}_\beta|$ are, apart from λ, S, k and $|\mathcal{C}_\beta|$, the quadrics of (Δ) whose cross-ratio with Γ, Σ and λ (taken twice), is rational or else they consist of sets of quadrics of (Δ) for which these cross-ratios are conjugate algebraic numbers. This last case

gives an example of an irreducible geometric variety which is a reducible invariant variety.

The variety $|\mathcal{C}_{\beta}|$ is an example of an irreducible invariant variety which is not of the form $|\mathcal{C}_{[a;\beta]}|$ for some element a of $\mathcal{C}_{,\beta}$ (see B 2.2g and B 2.2h), since every point of $|\mathcal{C}_{\beta}|$ lies on one of the quadrics of (Δ).

The tensor a_{β} possesses two independent simple invariants, namely $\psi = a \cdot \epsilon$ and $\chi = (a)^2 \cdot D^{(2;1;1)} \cdot (\epsilon)^2$. Each quadric of (Δ) is given by the absolute invariant ψ^2/χ taking a specified value. The quadric is a geometric variety when this value is rational.

B 3. Miscellaneous applications.

A number of applications of the preceding theory will be given, including a tensor proof of the basis theorem of Grassmann coordinates and the second fundamental theorem of invariant theory. Tensor expressions representing complexes of linear spaces and resultants will be given.

B 3.1 Some notations involving elements of $\mathfrak{J}$.

Let the elements $L^{(\lambda:\alpha)}$ and $M^{(\lambda,\alpha)}$ of $\mathfrak{J}$ be defined by

$$L^{(\lambda:\alpha)} = J_{[\alpha+1,\ldots,\alpha+\lambda,\,1,2,\ldots,\alpha]} \qquad \text{B 3.1a}$$

$$M^{(\lambda,\alpha)} = J_{[1,\lambda+1,\ldots,(\alpha-1)\lambda+1,\,2,\lambda+2,\ldots,(\alpha-1)\lambda+2,\;\ldots\ldots,\;\lambda,\,2\lambda,\ldots,\alpha\lambda]} \qquad \text{B 3.1b}$$

($J_{[\ldots]}$ is defined in A 1.5c and A 1.5h.) If a_{α^λ}, b_{α^λ}, c^{α^λ}, $d^{\alpha^\lambda} \in \mathcal{C}$, we have

$$ab \cdot L^{(\lambda:\alpha)} = ba,\; L^{(\alpha,\lambda)} \cdot cd = dc \;. \qquad \text{B 3.1c}$$

If $g_{(i)\beta}$, $h_{(i)}^{\beta} \in \mathcal{C}$ $(i = 1, .., r)$, then

$$\left.\begin{aligned} (g_{(1)})^{\Delta} \,..\, (g_{(\lambda)})^{\Delta} \bullet M_{(\beta)}^{(\Delta,\lambda)} &= \{g_{(1)} \,..\, g_{(\lambda)}\}^{\Delta}, \\ M_{(\beta)}^{(\lambda,\Delta)} \bullet (h_{(1)})^{\Delta} \,..\, (h_{(\lambda)})^{\Delta} &= \{h_{(1)} \,..\, h_{(\lambda)}\}^{\Delta}. \end{aligned}\right\} \quad \text{B } 3.1\text{d}$$

The elements $P^{(t,\lambda)}$, $N^{(t,\lambda)}$ and $M^{(t;\lambda,\Delta)}$ of $\mathfrak{S}$ are defined by

$$P_{(\beta)}^{(t,\lambda)} = P_{(\beta^t)}^{(\lambda)}, \quad N_{(\beta)}^{(t,\lambda)} = N_{(\beta^t)}^{(\lambda)}, \quad M_{(\beta)}^{(t;\lambda,\Delta)} = M_{(\beta^t)}^{(\lambda,\Delta)}. \qquad \text{B } 3.1\text{e}$$

We have $P^{(1,\lambda)} = P^{(\lambda)}$, $N^{(1,\lambda)} = N^{(\lambda)}$, $M^{(1;\lambda,\Delta)} = M^{(\lambda,\Delta)}$. Also,

$$P^{(t,\lambda)*} = P^{(t,\lambda)}, \quad N^{(t,\lambda)*} = N^{(\lambda)}, \quad M^{(t;\lambda,\Delta)*} = M^{(t;\Delta,\lambda)}, \quad L^{(\lambda:\Delta)*} = L^{(\Delta:\lambda)} \qquad \text{B } 3.1\text{f}$$

B 3.1g Definition. Let $A = \sum_i \lambda_i J_i$ be an element of $\mathfrak{S}$. (The $J_i \in \mathfrak{S}$ are permutations.) The element $\tilde{A}$ of $\mathfrak{S}$ is then defined by $\tilde{A} = \sum_i \eta_i \lambda_i J_i$, where η_i is +1 or −1 according as J_i is an even or odd permutation.

B 3.1h The operation $A \to \tilde{A}$ will be seen to be an automorphism of $\mathfrak{S}$, that is, $\widetilde{A+B} = \tilde{A} + \tilde{B}$, $\widetilde{AB} = \tilde{A}\,\tilde{B}$ and $\widetilde{A \bullet B} = \tilde{A} \bullet \tilde{B}$. These properties follow from linearity and the fact that $J_i\, J_j$ and $J_i \bullet J_j$ are even permutations if and only if J_i and J_j are both even or both odd. We also have $\tilde{\tilde{A}} = A$ and $\langle \tilde{A} \rangle_S = \langle A \rangle_S$, $\tilde{A^*} = (\tilde{A})^*$, $\tilde{A^\dagger} = (\tilde{A})^\dagger$ (see A 1.8k, A 2.10i).

B 3.1i $$\tilde{P}^{(\lambda,\Delta)} = \begin{cases} N^{(\lambda,\Delta)} \text{ if } r \text{ is odd} \\ P^{(\lambda,\Delta)} \text{ if } r \text{ is even} \end{cases}, \quad \tilde{N}^{(\lambda,\Delta)} = \begin{cases} P^{(\lambda,\Delta)} \text{ if } r \text{ is odd} \\ N^{(\lambda,\Delta)} \text{ if } r \text{ is even} \end{cases},$$

B 3.1j $\tilde{K}^{\{\lambda\}} = \pm K^{\{\mu\}*}$ where μ is the partition conjugate to λ (see A 1.8h),

B 3.1k $$\tilde{L}^{(\lambda:\Delta)} = (-1)^{\lambda\Delta} L^{(\lambda:\Delta)}, \quad \tilde{M}^{(t;\lambda,\Delta)} = (-1)^{t\frac{\lambda(\lambda-1)\Delta(\Delta-1)}{4}} M^{(t;\lambda,\Delta)}.$$

CHAPTER 6

B 3.11 Remark. From B 3.1i we have $\widetilde{N}^{(m+1)} = P^{(m+1)}$. It follows that the operation $A \to \widetilde{A}$ on $\mathfrak{J}$ does not extend uniquely to elements of $\mathcal{J}$ since $N_{(\alpha)}^{(m+1)} = 0$ but $P_{(\alpha)}^{(m+1)} \neq 0$ (see A 1.7). It is possible, however, to consider this operation as giving an isomorphism of $\mathcal{J}$ onto a normal tensor system $\widetilde{\mathcal{J}}$. The system $\widetilde{\mathcal{J}}$ turns out not to be a proper system, its basic index type being of class [0, n] (see A 1.9a). This isomorphism may be used to establish the consistency of systems possessing index types of class [0, n].

B 3.1m Notation. Let the element $K^{(\lambda,\Delta)}$ of $\mathfrak{J}$ be defined by

$$K^{(\lambda,\Delta)} = (N^{(\lambda)})^{\Delta} \bullet M^{(\lambda,\Delta)} \bullet (P^{(\Delta)})^{\lambda},$$

so that also, $K^{(\lambda,\Delta)*} = (P^{(\Delta)})^{\lambda} \bullet M^{(\Delta,\lambda)} \bullet (N^{(\lambda)})^{\Delta}$ (see B 3.1f). Let $p = N^{(\lambda)} \bullet x_{(1)} \ldots x_{(\lambda)}$ be an α-Grassmann tensor and $q = v_{(1)} \ldots v_{(\lambda)} \bullet N^{(\lambda)}$ be a dual α-Grassmann tensor ($x_{(1)}^{\alpha}, \ldots, x_{(\lambda)}^{\alpha}, v_{(1)\alpha}, \ldots, v_{(\lambda)\alpha} \in \mathcal{T}$, $0 \leqslant r \leqslant n$).

Then, by B 3.1d and B 1.2b,

$$K^{(\lambda,\Delta)} \bullet (x_{(1)})^{\Delta} \ldots (x_{(\lambda)})^{\Delta} = (p)^{\Delta}, \quad (v_{(1)})^{\Delta} \ldots (v_{(1)})^{\Delta} \bullet K^{(\lambda,\Delta)*} = (q)^{\Delta}. \qquad \text{B 3.1n}$$

It follows from this that $K^{(\lambda,\Delta)} \neq 0$ if $r \leqslant n$. Also, the condition $r \leqslant n$ is irrelevant for elements of $\mathfrak{J}$, since $\mathfrak{J}$ is defined without reference to n. Hence, $K^{(\lambda,\Delta)} = \pm K^{\{\Delta^{\lambda}\}}$ (see A 1.8g and A 1.8h). Put $G^{(\lambda,\Delta)} = G^{\{\Delta^{\lambda}\}}$ Then by A 1.8j and by A 1.8i and A 2.2o, we have

$$G^{(\lambda,\Delta)} = \kappa_{(\Delta^{\lambda})} K^{(\lambda,\Delta)} \bullet K^{(\lambda,\Delta)*} = K^{(\lambda,\Delta)} \bullet K^{(\lambda,\Delta)\dagger} = K^{(\lambda,\Delta)*\dagger} \bullet K^{(\lambda,\Delta)*}. \qquad \text{B 3.1o}$$

Also, using B 3.1j and B 3.1h, the element $H^{(\lambda,\delta)}$ of $\mathfrak{J}$ can be defined by

$$H^{(\lambda,\delta)} = \tilde{G}^{(\lambda,\delta)} = K^{(\lambda,\delta)*} \bullet K^{(\lambda,\delta)*\dagger} = K^{(\lambda,\delta)\dagger} \bullet K^{(\lambda,\delta)}. \qquad \text{B 3.1p}$$

B 3.2 Theorem. Let $p^{\alpha\lambda}$ and $q_{\alpha\lambda}$ be, respectively, an α-Grassmann tensor and a dual α-Grassmann tensor. Then $E^{(\delta)}_{\{p;\mathfrak{R}\}} = E^{(\delta)}_{\{q;\mathfrak{R}\}} = E^{(\delta)}_{\{p\}} = E^{(\delta)}_{\{q\}} = G^{(\lambda,\delta)}$.

Proof. We have $q = v_{(1)} \,..\, v_{(\lambda)} \bullet N^{(\lambda)}$; where $v_{(1)}$, .., $v_{(\lambda)}$ are linearly independent. Let $A^{\alpha\lambda}_{\beta} \in \mathfrak{R}$. Then, by B 3.1n, $(q)^{\delta} \bullet A = 0$ if and only if $(v_{(1)})^{\delta} \,..\, (v_{(\lambda)})^{\delta} \bullet K^{(\lambda,\delta)*} \bullet A = 0$, that is (by B 2.2), if and only if $E^{(\delta,\ldots,\delta)}_{\{v_{(1)},\ldots,v_{(\lambda)}\}} \bullet K^{(\lambda,\delta)*} \bullet A = 0$. Hence, by B 2.2a, $(q)^{\delta} \bullet A = 0$ if and only if $(P^{(\delta)})^{\lambda} \bullet K^{(\lambda,\delta)*} \bullet A = 0$, that is, $K^{(\lambda,\delta)*} \bullet A = 0$. Now, by B 3.1o, $G^{(\lambda,\delta)} = K^{(\lambda,\delta)*\dagger} \bullet K^{(\lambda,\delta)*}$ $\sim\!K\!\sim\!\sim\!K\!\sim$, so that, by A 4.1a, $G^{(\lambda,\delta)}$ is a projector in $\mathfrak{R}$ with the same right annihilators in $\mathfrak{R}$ as $(q)^{\delta}$ has. Thus $G^{(\lambda,\delta)} = E^{(\delta)}_{\{q\}}$. The proofs for the other projectors are similar.

B 3.2a Definition. Let the Geometric varieties of α-Grassmann tensors $p^{\alpha\lambda}$ and of dual α-Grassmann tensors $q_{\alpha\lambda}$ be denoted, respectively, by $\mathcal{U}^{G}_{(\nu,\lambda)}$ and $\mathcal{U}_{G}(\lambda)$. The varieties $|\mathcal{U}^{G}(\lambda)|$ and $|\mathcal{U}_{G}(\lambda)|$ are Grassmann varieties.

B 3.2b Corollary. $E^{(\delta)}_{\mathcal{U}^{G}(\nu,\lambda)} = E^{(\delta)}_{\mathcal{U}_{G}(\lambda)} = G^{(\lambda,\delta)}$.

Proof. This follows from B 3.2 and B 2.2s. Hence, by B 2.3 and B 2.3a, we have

B 3.2c Corollary. A necessary and sufficient condition for $a \bullet (p)^{\lambda} = 0$ for all $p \in \mathcal{U}^{G}_{(\nu,\lambda)}$ is $a \bullet G^{(\lambda,\delta)} = 0$. Also, $G^{(\lambda,\delta)} \bullet (p)^{\delta} = (p)^{\delta}$ for all $p \in \mathcal{U}^{G}_{(\nu,\lambda)}$. A similar result holds for $\mathcal{U}_{G}(\lambda)$.

B 3.2d This result gives a canonical form for polynomials in Grassmann coordinates (see B 2.3b and B 2.3c) since the form $b\bullet(p)^{\Delta}$ in the Grassmann tensor p uniquely determines b if $b = b\bullet G^{(\lambda,\Delta)}$.

B 3.2e Corollary. The postulation formula for the Grassmann variety $\mathcal{U}^{G}_{(\lambda)}$ of (r-1)-spaces in an (n-1)-space, for primals of order s, is

$$\langle G^{(\lambda,\Delta)} \rangle = \begin{vmatrix} \binom{m+\Delta-1}{\Delta} & \binom{m+\Delta}{\Delta+1} & \cdots & \binom{m+\Delta+\lambda-2}{\Delta+\lambda-1} \\ \binom{m+\Delta-2}{\Delta-1} & \binom{m+\Delta-1}{\Delta} & & \vdots \\ \vdots & & \ddots & \vdots \\ \binom{m+\Delta-\lambda}{\Delta-\lambda+1} & \cdots & \cdots & \binom{m+\Delta-1}{\Delta} \end{vmatrix} = \prod_{i=1}^{\Delta} \prod_{j=1}^{\lambda} \frac{(m+i-j)}{(i+j-1)} \quad .$$

Proof. This follows from B 2.4i, B 3.2b and A 1.8y (see also (10)p371,... and (8) for the determinantal formula. The product formula is due to D.E. Littlewood)

B 3.2f A particular case of B 3.2 is geven when $p = \epsilon$ and $q = \epsilon^{*}$ giving $E^{(\Delta)}_{\{\epsilon\}} = E^{(\Delta)}_{\{\epsilon^{\Delta}\}} = G^{(m,\Delta)}_{(\alpha)}$. But since, by A 4.5b, $\epsilon\epsilon^{*} = n!N^{(m)}_{(\alpha)}$ it follows that $E^{(\Delta)}_{\{\epsilon\}} = (N^{(m)}_{(\alpha)})^{\Delta}$. Hence $G^{(m,\Delta)}_{(\alpha)} = (N^{(m)}_{(\alpha)})^{\Delta}$. However, it is not true that $G^{(m,\Delta)} = (N^{(m)})^{\Delta}$ for elements of $\mathfrak{J}$ (if $s > 1$). The correct formula (see A 1.9b and A 1.9c) is given by

$$\{G^{(m,\Delta)} - (N^{(m)})^{\Delta}\}\bullet\{I^{(m\Delta)} - Q^{(m\Delta)}_{m}\} = 0.$$

B 3.3 Theorem. Let a_{α_1,α_2} be an α-skew-symmetric element of $\mathcal{C}_{\alpha^2}$. Then a possesses no simple absolute invariants other than constants.

Proof. A simple absolute invariant λ of a is an element of $\mathfrak{R}$ satisfying $\Psi_{(1)} = \lambda\Psi_{(2)}$ where $\Psi_{(1)}$ and $\Psi_{(2)}$ are simple invariants of a of equal weight r and $\Psi_{(2)}$ is not identically zero (see B 1.3c).

Let $\Psi_{(1)} = (a)^{\Delta}\bullet A\bullet(\ \epsilon\ )^{\lambda}$ and $\Psi_{(2)} = (a)^{\Delta}\bullet B\bullet(\ \epsilon\ )^{\lambda} \neq 0$, where $A, B \in \mathfrak{J}$ and $2s = nr$. (The tensor a, being covariant, clearly has no

invariants of negative weight.) It is required to prove that, if

$$(a)^{d}\cdot A\cdot(\ \in\ )^{\lambda} = \lambda\,(a)^{d}\cdot B\cdot(\ \in\ )^{\lambda},$$

then λ is independent of a. That is, it is required to show that, for some λ of $\mathfrak{R}$, the identity

$$P^{(\lambda,d)}\cdot(N^{(\lambda)})^{d}\cdot A\cdot G^{(m,\lambda)} = \lambda\ P^{(\lambda,d)}\cdot(N^{(\lambda)})^{d}\cdot B\cdot G^{(m,\lambda)}$$

holds for elements of $\mathcal{F}$ by B 3.2c (see B 3.2f) and since the general skew-symmetric tensor $a_{\alpha_1\alpha_2}$ is the general tensor of the form $c\cdot N^{(2)}_{(\alpha)}$ ($c_{\alpha_1\alpha_2}\in\mathcal{C}$). It will be sufficient to show that this identity is true for elements of $\mathfrak{F}$ (for some λ). By $A\to\tilde{A}$ (see B 3.1g,..,B 3.1j) this is equivalent to the identity

$$P^{(\lambda,d)}\cdot(P^{(\lambda)})^{d}\cdot A\cdot K^{(n,m)\dagger}\cdot K^{(n,m)} = \lambda\,P^{(\lambda,d)}\cdot(P^{(\lambda)})^{d}\cdot\tilde{B}\cdot K^{(n,m)\dagger}\cdot K^{(n,m)}$$

by B 2.2i and B 2.2n.

Now $K^{(n,m)} = (N^{(n)})^{m}\cdot M^{(n,m)}\cdot(P^{(m)})^{n}$ so that it can be written in the form $C\cdot\{(P^{(m)})^{n}\} = 0$. Thus lemma A 1.9d applies, showing that the identity is true for elements of $\mathfrak{F}$ if it is true for elements of the system $\mathcal{F}(\mathfrak{S}')$ where $\mathfrak{S}'$ is semigenerated by a single element σ of dimension r. But $\mathcal{F}(\mathfrak{S}')$ is a subsystem of $\mathcal{C}' = \mathcal{C}(\mathfrak{S}',\mathfrak{R})$. Thus $\mathcal{C}'$ contains a σ-alternating tensor $\epsilon^{(\sigma)}$ and symmetric elements $b_{\sigma_1\sigma_2}$. Also the results of section B can be applied with r in place of n.

Now, since $G^{(n,m)} = K^{(n,m)}\cdot K^{(n,m)\dagger}$, the $K^{(n,m)}$ of the identity can be replaced by $G^{(n,m)}$. It is thus equivalent to the statement that, for all symmetric $b_{\sigma_1\sigma_2}$ of $\mathcal{C}'$, the relation

CHAPTER 6

$$(b)^{\Delta}\bullet\tilde{A}\bullet K^{(\lambda,n)\dagger}\bullet(\epsilon^{(\sigma)})^{n} = \lambda(b)^{\Delta}\bullet\tilde{B}\bullet K^{(\lambda,n)\dagger}\bullet(\epsilon^{(\sigma)})^{n}$$

holds for fixed λ (by B 3.2c). The left hand side of this is not identically zero, for otherwise this would imply $\Psi_{(\Delta)} = 0$. Also, both sides are scalars so that this defines λ. But this definition makes λ an absolute invariant of b, being a quotient of two relative invariants of weight n. Now the invariants of b are those of the quadratic form $b\bullet(x)^2$ since $b_{\sigma_1\sigma_2}$ is symmetric. However, it is well-known that a quadratic form possesses no absolute invariants other than constants. This proves the result.

B 3.3a Corollary. Every skew-symmetric determinant of even order is a perfect square.

Proof. Let n be even and $a_{\alpha_1\alpha_2} \in \mathcal{C}_{\alpha^2}$. Now the determinant of the matrix of components of a is given by $\frac{1}{n!}(a)^{n}\bullet M^{(2,n)}\bullet(\epsilon)^{2}$. (See Eisenhart (5) pp. 101-102.)

I shall show that if n is even and $a_{\alpha_1\alpha_2}$ is skew-symmetric, then

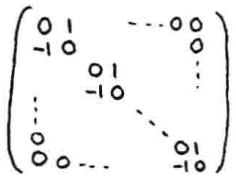
$$\tfrac{1}{n!}(a)^{n}\bullet M^{(2,n)}\bullet(\epsilon)^{2} = \left\{\frac{1}{(n/2)!\,2^{n/2}}(a)^{n/2}\bullet\epsilon\right\}^{2}.$$

This will give the required square root explicitly.

Define the scalar λ by

$$\lambda\left\{\tfrac{1}{n!}(a)^{n}\bullet M^{(2,n)}\bullet(\epsilon)^{2}\right\} = \left\{(a)^{n/2}\bullet\epsilon\right\}^{2}.$$

Then λ is a simple absolute invariant of a. Thus, by B 3.3, it is a constant. It is then easily verified that $\lambda = \{(\frac{n}{2})!\}^{2}2^{n}$, by considering the case when a has components forming the matrix



for then $(a)^{n/2} \cdot \epsilon = \left(\frac{n}{2}\right)! \, 2^{n/2}$.

(For an alternative proof, see Turnbull (21) PP 105-107.)

B 3.4 <u>The number of linearly independent covariants of type $\{\lambda\}$ of a given set of tensors.</u>

Let a_{α^*}, .., $c_{\alpha^\wedge}$ be given elements of $\mathcal{C}$. Let λ be a partition of $m = ps + \ldots + ru$, where s, .., u are non-negative integers and let v be a Clebsch variable tensor of type $\{\lambda\}$. It is required to find, for the given a, .., c, the number of linearly independent covariants of the form

$$(a)^{\alpha} \ldots (c)^{\mu} \cdot X \cdot v^{\{\lambda\}} , \qquad \text{B } 3.4a$$

where $X \in \mathcal{T}_{\alpha^m \alpha^{*m}}$. (See B 1.3b). Such covatiants clearly form a linear set over $\mathfrak{R}$. Also the covariant B 3.4a vanishes identically if and only if $(a)^{\alpha} \ldots (c)^{\mu} \cdot X \cdot K^{\{\lambda\}} = 0$. Thus, it is required to find the number of linearly independent tensors,

$$(a)^{\alpha} \ldots (c)^{\mu} \cdot X \cdot K^{\{\lambda\}} .$$

Putting $\mathcal{Q} = \mathcal{Q}_{\{a,\ldots,c;\mathfrak{R}\}}$. Then by B 2.2 and B 2.2p, this is the number of linearly independent tensors of the form

$$E^{(\alpha,\ldots,\omega)}_{\mathcal{Q}} \cdot X \cdot K^{\{\lambda\}} \qquad \text{B } 3.4b$$

for which $X \in \mathcal{T}_{\alpha^m \alpha^{*m}}$.

Now $K^{\{\lambda\}} = N^{(\mu_1)} \ldots N^{(\mu_{\ell'})} \cdot J^{\{\lambda\}} \cdot P^{(\lambda_1)} \ldots P^{(\lambda_\ell)}$, where $\mu = (\mu_1 \ldots \mu_{\ell'})$ is the partition of m conjugate to $\lambda = (\lambda_1 \ldots \lambda_\ell)$. Hence lemma A 1.9d applies provided that $n \geqslant 1 = \max(\mu_1, \ldots, \mu_{\ell'}) = \mu_1$, so that, in that case, the number of linearly independent tensors B 3.4b is equal

to the number of linearly independent elements of $\mathfrak{F}$ which have the form B 3.4b. The symbol $K^{\{\lambda\}}$ represents ~~to~~ a well-defined element of $\mathfrak{F}$, while the symbol $E_{\mathfrak{U}}^{(a,\cdots,\mu)}$ can be taken to represent any element of $\mathfrak{F}$ which corresponds to the tensor $E_{\mathfrak{U}}^{(a,\cdots,\mu)}$. More definitively we may choose $E_{\mathfrak{U}}^{(a,\cdots,\mu)}$ to represent that element of $\mathfrak{F}$ for which $E_{\mathfrak{U}}^{(a,\cdots,\mu)} \cdot Q_m^{(m)} = 0$. (See A 1.9c).

If $n < 1$, then, since $K^{\{\lambda\}}$ involves $N^{(\ell)} = N^{(\mu_1)}$, it follows that all expressions of the form B 3.4b represent a zero element of $\mathfrak{F}$. Taking $E_{\mathfrak{U}}^{(a,\cdots,\mu)}$ to satisfy $E_{\mathfrak{U}}^{(a,\cdots,\mu)} \cdot Q_m^{(m)} = 0$, B 3.4b then also represents a zero of $\mathfrak{F}$.

Thus, in any case it is the number of linearly independent elements of $\mathfrak{F}$ of the form B 3.4b that is required. B A 2.11a, this is equal to

$$\frac{1}{m!} \sum_{i=1}^{m!} \left\langle E_{\mathfrak{U}}^{(a,\cdots,\mu)} \cdot J_i \cdot G^{\{\lambda\}} \cdot J_i^* \right\rangle_S \qquad \text{B 3.4c}$$

since $G^{\{\lambda\}} = K^{\{\lambda\}} \cdot K^{\{\lambda\}\dagger}$ and $E_{\mathfrak{U}}^{(a,\cdots,\mu)}$ is a projector. But, by A 1.8q we have

$$\frac{\chi_i^{(\lambda)}}{m!} \sum_{i=1}^{m!} J_i \cdot G^{\{\lambda\}} \cdot J_i^* = T^{\{\lambda\}},$$

so that the expression B 3.4c is equal to

$$\frac{1}{\chi_1^{(\lambda)}} \left\langle E_{\mathfrak{U}}^{(a,\cdots,\mu)} \cdot T^{\{\lambda\}} \right\rangle_S . \qquad \text{B 3.4d}$$

Let $E_{\mathfrak{U}}^{(a,\cdots,\mu)} = \sum_{i=1}^{m!} \xi_i J_i$. Now, by A 1.8f, we have $T^{\{\lambda\}} = \frac{\chi_1^{(\lambda)}}{m!} \sum_{i=1}^{m!} \chi_i^{(\lambda)} J_i^*$ so that, by A 1.8m, the expression B 3.4d is equal to

$$\sum_{i=1}^{m!} \xi_i \chi_i^{(\lambda)} . \qquad \text{B 3.4e}$$

B 3.4f By A1.8w, A1.8x, the expression B 3.4d is also the coefficient of the Schur function $\{\lambda\}$ in the symmetric function $\{E_{\mathfrak{U}}^{(a,\cdots,\mu)}\}_S$.

These results show that the number of linearly independent forms B 3.4a can be calculated from a knowledge of the tensor $E^{(a,\ldots,w)}_{\{a,\ldots,c;\lambda\}} = E^{(a,\ldots,w)}_{\mathcal{U}}$, this number being given by B 3.4c,.., B 3.4f.

B 3.4g The preceding arguments can also be applied to the problem of finding the number of covariants B 3.4a which are linearly independent functions for all (a, .., c) belonging to some invariant variety, by denoting this variety by $\mathcal{U}$ instead of putting $\mathcal{U} = \mathcal{V}_{\{a,\ldots,c;\lambda\}}$. The number of covariants B 3.4a which are linearly independent functions for all (a, .., c) $\in \mathcal{U}$ is then equal to the number of linearly independent expressions B 3.4b, by B 2.3, and this number is again given by B 3.4c, .., B 3.4f.

B 3.4h In particular, if $\mathcal{U}$ is the set of all (a, .., c) where $a_{\alpha_1 \ldots \alpha_p}$, ..., $c_{\alpha_1 \ldots \alpha_\lambda}$ are <u>symmetric</u> elements of $\mathcal{C}$ (i.e. all (a, .., c) for which $a \cdot P^{(p)} = a$, .., $c \cdot P^{(\lambda)} = c$), then the number of expressions B 3.4a which are linearly independent for all (a, .., c) of $\mathcal{U}$ is the number of covariants of type $\{\lambda\}$ of respective degrees s, .., u in forms of orders p, .., r, which are linearly independent for all such forms.

In this case we have $E^{(a,\ldots,w)}_{\mathcal{U}} = \{P^{(p,q)} \cdot (P^{(p)})^q\} \ldots \{P^{(\lambda,\omega)} \cdot (P^{(\lambda)})^\omega\}$.

Thus, by A 1.8w, we have

$$\{E^{(a,\ldots,w)}_{\mathcal{U}}\}_S = \{P^{(p,q)} \cdot (P^{(p)})^q\}_S \ldots \{P^{(\lambda,\omega)} \cdot (P^{(\lambda)})^\omega\}_S$$
$$= (\{p\} \otimes \{s\}) \ldots (\{r\} \otimes \{u\}).$$

Applying this to B 3.4f, we obtain the usual Scmur functional method of obtaining the number and type of covariants of a given set of forms. (See

Littlewood (11) pp288-9).

With this definition for $\mathcal{U}$, it is clearly an irreducible geometric variety. Thus (a, .., c) may be chosen to be a generic set of tensors of $\mathcal{U}$. (See B 2.2g and B 2.2i.) Then $\mathcal{U} = \mathcal{U}_{\{a,..,c\}}$, so that $E_{\mathcal{U}}^{(a,..,\omega)} = E_{\{a,..,c\}}^{(a,..,\omega)}$.

It follows that B 3.4c, .., B 3.4f give the number of expressions B 3.4a for which $X \in \mathfrak{R}$, which are linearly independent over the rationals.

This is not necessarily the same as the number which are linearly independent over $\mathfrak{K}$. For example, suppose a, .., c possess two invariants ψ_1 and ψ_2, of equal weight w where

$$\psi_1 = (a)^{\alpha} .. (c)^{\mu} \cdot X \cdot (\in)^{\omega} \text{ and } 0 \neq \psi_2 = (a)^{\alpha} .. (c)^{\mu} \cdot Y \cdot (\in)^{\omega},$$

X and Y being elements of $\mathfrak{R}$. If ψ_1 and ψ_2 are not formally proportional, then the ratio ψ_1/ψ_2 will be an irrational element of $\mathfrak{K}$, (a simple absolute invariant of a, .., c). Thus, the invariants ψ_1 and ψ_2 will be linearly independent over the rationals but linearly dependent over $\mathfrak{K}$ (trivially, since $\psi_1 - (\psi_1/\psi_2)\psi_2 = 0$). Of course ψ_1 and ψ_2 are not linearly dependent functions of the tensors a, .., c. This is the situation treated in B 3.4g.

B 3.5 Complexes of linear spaces.

Let $p^{\alpha_1 .. \alpha_r}$ and $q_{\alpha_1 .. \alpha_{m-r}}$ be respectively, an α-Grassmann tensor and a dual α-Grassmann tensor corresponding to an r-1 dimensional linear space,

p and q being related by the formulae of A 4.5d ; that is, $p = \frac{1}{\lambda!} q I^{(\lambda)} \cdot \epsilon$ and $q = \frac{1}{(n-\lambda)!} \epsilon^* \cdot I^{(n-\lambda)} p$.

Let $\psi \equiv \psi(\ldots, p^{i_1 \ldots i_\lambda}, \ldots) \equiv \psi(p)$ be a homogeneous polynomial of degree s in the components $p^{i_1 \ldots i_\lambda}$ (in the frame $\mathfrak{F}$) of the tensor p. Then, by B 3.2d, ψ is of the form

$$\psi(p) \equiv f_G \cdot (p)^A ,$$

where $f_{G\,\kappa_1 \ldots \kappa_{sA}}$ is some element of $\mathcal{C}_{\alpha rs}$ and where, without loss of generality,

$$f_G = f_G \cdot G^{(\lambda,A)} . \qquad \text{B 3.5a}$$

Also, the function $\psi(p)$ and this tensor f_G determine each other uniquely.

Now, by A 4.5g, $f_G \cdot (p)^A = \frac{1}{(\lambda!)^A} (q)^A \cdot \{ I^{((n-\lambda)A)} f_G \cdot D^{(n-\lambda : \lambda ; A)} \cdot (\epsilon)^A \}$, so that

$$\psi \equiv (q)^A \cdot f^G$$

where f^G is an element of $\mathcal{C}_{\alpha * (n-\lambda)A}$ and, without loss of generality,

$$f^G = G^{(n-\lambda,A)} \cdot f^G . \qquad \text{B 3.5b}$$

We have $f^G = \frac{1}{(\lambda!)^A} G^{(n-\lambda,A)} f_G \cdot D^{(n-\lambda : \lambda ; A)} \cdot (\epsilon)^A$ by B 3.2d. In fact

$$f^G = \frac{1}{(\lambda!)^A} I^{((n-\lambda)A)} f_G \cdot D^{(n-\lambda : \lambda ; A)} \cdot (\epsilon)^A . \qquad \text{B 3.5c}$$

This follows from B 3.2c (see B 3.2f), B 3.5a and the following lemma :

B 3.5d **Lemma.** $I^{((n-\lambda)A)}\, G^{(\lambda,A)} \cdot D^{(n-\lambda : \lambda ; A)} \cdot G^{(n,A)} = G^{(n-\lambda,A)} G^{(\lambda,A)} \cdot D^{(n-\lambda : \lambda ; A)} \cdot G^{(n,A)}$

Proof. Let $a^{\alpha m} \in \mathcal{U}^G_{(n-m)}$ and $b_{\alpha\lambda} \in \mathcal{U}_{G(\lambda)}$. Then, by A 4.4h (and A 4.4g), the intersection of the spaces a and b is given by the tensor $I^{(n-\lambda)} b \cdot a \in \mathcal{U}^G_{(n-\lambda+\mu)}$ unless $I^{(n-\lambda)} b \cdot a = 0$. Hence, by B 3.2c,

$\{ I^{((n-\lambda)A)} - G^{(n-\lambda,A)} \} (I^{(n-\lambda)} b \cdot a)^A = 0$, that is, $\{ I^{((n-\lambda)A)} - G^{(n-\lambda,A)} \} (b)^A \cdot D^{(n-\lambda : \lambda ; A)} \cdot (a)^A = 0$.

The result follows from B 3.2c.

Similarly to B 3.5c we get

$$f_{\mathfrak{G}} = \frac{1}{\{(n-\lambda)!\}^{\mu}} (\epsilon^*)^{\Delta} \bullet D^{(n-\lambda;\lambda;a)*} \bullet f^{\mathfrak{G}} I^{(\lambda\mu)}. \qquad \text{B 3.5e}$$

Now $p = N^{(\lambda)} \bullet x_{(1)} \dots x_{(\lambda)}$ and $q = u_{(1)} \dots u_{(n-\lambda)} \bullet N^{(n-\lambda)}$ where $x^{\alpha}_{(1)}, \dots, x^{\alpha}_{(\lambda)}$ and $u_{(1)\alpha}, \dots, u_{(n-\lambda)\alpha}$ are linearly independent elements of $\mathfrak{C}_{\alpha^*}$ and of $\mathfrak{C}_{\alpha}$, respectively. Thus, by B 3.1n,

$$\Psi \equiv f_H \bullet (x_{(1)})^{\Delta} \dots (x_{(\lambda)})^{\Delta} \equiv (u_{(1)})^{\Delta} \dots (u_{(n-\lambda)})^{\Delta} \bullet f^H$$

where

$$f_H = f_{\mathfrak{G}} \bullet K^{(\lambda,\Delta)} \text{ and } f^H = K^{(\mu\Delta,\Delta)*} \bullet f^{\mathfrak{G}}. \qquad \text{B 3.5f}$$

By B 3.1p, B 3.5a and B 3.5b we have

$$f_{\mathfrak{G}} = f_H \bullet K^{(\lambda,\mu\Delta)\dagger} \text{ and } f^{\mathfrak{G}} = K^{(n-\lambda,\Delta)*\dagger} \bullet f^H \qquad \text{B 3.5g}$$

whence

$$f_H = f_H \bullet H^{(\Delta,\lambda)} \text{ and } f^H = H^{(\Delta,n-\lambda)} \bullet f \,. \qquad \text{B 3.5h}$$

B 3.5i Any one of $f_{\mathfrak{G}}$, $f^{\mathfrak{G}}$, f_H, f^H can be obtained from each of the others by means of B 3.5c, B 3.5e, B 3.5f, B 3.5g. I shall use such a notation (e.g. $a_{\mathfrak{G}}$, $a^{\mathfrak{G}}$, a_H, a^H) only for tensors connected by such relations. The set $f_{\mathfrak{G}}$, $f^{\mathfrak{G}}$, f_H, f^H is determined by the element $f_{\mathfrak{G}} \neq 0$ of $\mathfrak{C}_{\alpha^{\mu}}$, subject only to B 3.5a. Similarly, it follows that the only restriction on $f_H \neq 0$ as an element of $\mathfrak{C}_{\alpha^{\mu}}$ is that $f_H \bullet H^{(\Delta,\lambda)} = f_H$ as in B 3.5h.

B 3.5j By B 2.4e and B 3.2b, we have $G^{(\Delta,\lambda)} = G^{(\Delta,\lambda)} \bullet P^{(\Delta,\lambda)}$. Hence, by B 3.1h and B 3.1p, $H^{(\Delta,\lambda)} = H^{(\Delta,\lambda)} \bullet \widetilde{P}^{(\Delta,\lambda)}$. Also $H^{(\Delta,\lambda)} = H^{(\Delta,\lambda)} \bullet P^{(\Delta)} I^{(\lambda\mu-\Delta)}$ since $H^{(\Delta,\lambda)} = K^{(\lambda,\Delta)\dagger} \bullet (N^{(\lambda)})^{\Delta} \bullet M^{(\Delta,\Delta)} \bullet (P^{(\Delta)})^{\lambda}$ (see B 3.1m). Thus f_H has the properties

$$f_H = f_H \bullet \widetilde{P}^{(\Delta,\lambda)} \text{ and } f_H = f_H \bullet P^{(\Delta)} I^{(\lambda\mu-\Delta)}.$$

This first relation shows that the function $f_H \bullet (x_{(1)})^s \ldots (x_{(r)})^s$ is symmetric or skew-symmetric in $x_{(1)}, \ldots, x_{(r)}$ according as s is even or odd (see B 3.11i). The second relation shows that f_H is symmetric in its first s indices and hence, because of the first relation, symmetric in each of its r groups of s indices. (This is in accordance with B 1.2c.)

B 3.5k The function $\psi \equiv f_H \bullet (x_{(1)})^s \ldots (x_{(r)})^s$ has the property that if $x_{(1)}, \ldots, x_{(r)}$ are linearly independent, the condition $\psi = 0$ is a property only of the (r-1)-space joining $x_{(1)}, \ldots, x_{(r)}$. Now, if $x_{(2)}$ is replaced by $x_{(2)} + \lambda x_{(1)}$ with $\lambda \neq 0$, we should have $f_H \bullet (x_{(1)})^s (x_{(2)} + \lambda x_{(1)})^s (x_{(3)})^s \ldots (x_{(r)})^s = 0$ whenever $\psi = 0$, that is

$$f_H \bullet (x_{(1)})^s \ldots (x_{(r)})^s + \lambda \sum_{i=1}^{s} f_H \bullet (x_{(1)})^{s-i} (x_{(2)})^{i-1} x_{(1)} (x_{(2)})^{s-i} (x_{(3)})^s \ldots (x_{(r)})^s + \lambda^2 \{\ldots\} + \ldots = 0.$$

This expression and ψ are both polynomials in the components of $x_{(1)}, \ldots, x_{(r)}$, in some frame, and have the same zeros. Thus they are proportional and the coefficient of λ in the above expression must be identically zero. By the symmetry of f_H, this condition is $f_H \bullet (x_{(1)})^{s-1} (x_{(2)})^{s-1} (x_{(3)})^s \ldots (x_{(r)})^s \equiv 0$, whence, by B 1.2, and the symmetry of f_H,

$$f_H \bullet P^{(s+1)} I^{(rs-s-1)} = 0.$$

B 3.51 Conversely, if $X \equiv a \bullet (x_{(1)})^s \ldots (x_{(r)})^s$, where $a \in \mathcal{C}_{\alpha^{rs}}$, and a satisfies

$$a = a \bullet \widetilde{P}^{(s,r)}, \quad a = a \bullet P^{(s)} I^{(s-1)}, \quad a \bullet P^{(s+1)} I^{(rs-s-1)} = 0,$$

then X has the property that the condition $X = 0$ depends only on the

CHAPTER 6

(r-1)-space joining $x_{(1)},..,\ x_{(\lambda)}$ ($x_{(1)},..,x_{(\lambda)}$ being linearly independent). For, by A 1.5r, $a\bullet P^{(a+1)}I^{(\lambda a-a-1)} = 0$ implies $a\bullet P^{(a+1)}I^{(\lambda a-a-2)} = .. = a\bullet P^{(2a)}I^{(\lambda a-2a)} = 0$, so that, reversing the argument of B 3.5k, we have $X = a\bullet(x_{(1)})^a(x_{(2)}+\lambda\, x_{(1)})^a(x_{(3)})^a..\ (x_{(\lambda)})^a$. By the [skew-]symmetry of a this holds for any pair $x_{(i)}$, $x_{(j)}$ and the result follows.

B 3.5m The condition $z\bullet(x_{(1)})^{a+1}(x_{(2)})^{a-1}(x_{(3)})^a..\ (x_{(\lambda)})^a = 0$ may be written $\sum_{\delta=1}^{n} x_{(1)}^{\delta}\frac{\partial X}{\partial x_{(2)}^{\delta}} = 0$ where $x_{(1)}^{\delta}$ and $x_{(2)}^{\delta}$ are the components, in some frame, of $x_{(1)}$ and $x_{(2)}$. It is a theorem due to Hodge (see (8) and also (10) pp.41-45) that if this condition holds for each pair of variables $x_{(1)},..,x_{(\lambda)}$, then X is a homogeneous polynomial of degree s in the Grassmann coordinates of the (r-1)-space joining $x_{(1)},..,\ x_{(\lambda)}$. If it is assumed that X is [skew-] symmetric (and this is not hard to prove from the given condition), then in virtue of B 3.5l, this result is a consequence of the following lemma :

B 3.5n <u>Lemma</u>. If $X \equiv a\bullet(x_{(1)})^a..\ (x_{(\lambda)})^a$, $a_{\alpha^{\lambda a}} \in \mathcal{C}$ and the condition $X = 0$ is a property only of the (r-1)-space joining $x_{(1)},..,\ x_{(\lambda)}$, provided they are linearly independent, then $X \equiv a\bullet K^{(a,\lambda)\dagger}\bullet(p)^\lambda$, where $p = N^{(\lambda)}\bullet x_{(1)}..\ x_{(\lambda)}$.

<u>Proof</u>. Choose a fixed space p and consider a projection Ψ of $\mathcal{C}$ with base space p (see B 1.5). Let e be the idempotent defining Ψ. Then, by B 1.5c, $(e)^\lambda\bullet p = p$, so that Ψ is an idempotent linear transformation of the space p. It thus sends each point of this space into itself, so that $e\bullet x_{(i)} = x_{(i)}$. The function $\Psi(X) = \{\Psi(a)\}\bullet(x_{(1)})^a..(x_{(\lambda)})^a$; $x_{(i)} \in \Psi(\mathcal{C}_{\alpha^a})$, is either identically zero or else it can vanish only when $x_{(1)},\ ..,\ x_{(\lambda)}$ are linearly dependent. Hence, by the argument given in the proof of B 1.4, in either

case, $\{\Psi(a)\}\bullet(P^{(q)})^{\lambda} = K_{(q^{\lambda})}\{\Psi(a)\}\bullet K^{(q,\lambda)}\bullet K^{(q,\lambda)}$, so that, by B 3.1n, $\{\Psi(a)\}\bullet(x_{(1)})^{q}..(x_{(\lambda)})^{q} = K_{(q^{\lambda})}\{\Psi(a)\}\bullet K^{(q,\lambda)}\bullet(p)^{q}$. Now $\{\Psi(a)\}\bullet(x_{(1)})^{q}..(x_{(\lambda)})^{q}$ $= \{a\bullet(e)^{\lambda q}\}\bullet(x_{(1)})^{q}..(x_{(\lambda)})^{q} = a\bullet(x_{(1)})^{q}..(x_{(\lambda)})^{q}$ and, similarly, $\{\Psi(a)\}\bullet K^{(q,\lambda)}\bullet(p)^{q} = a\bullet K^{(q,\lambda)}\bullet(p)^{q}$, whence $a\bullet(x_{(1)})^{q}..(x_{(\lambda)})^{q} = a\bullet K^{(q,\lambda)}\bullet(p)^{q}$. This holds for each choice of p, proving the result.

The idea of proving Hodge's theorem by means of the first main theorem of invariant theory is due to van der Waerden (see Pedoe (16)).

B 3.5o Lemma. Let $a_{\alpha\mu} \in \mathcal{C}$. Then $a = a\bullet H^{(q,\lambda)}$ if and only if $a = a\bullet\widetilde{P}^{(q,\lambda)}$, $a = a\bullet P^{(q)}I^{(\mu-\lambda)}$ and $a\bullet P^{(q+1)}I^{(\mu-\lambda-1)} = 0$.

Proof. B 3.5j and B 3.5k show that $a = a\bullet H^{(q,\lambda)}$ implies the other relations (see B 3.5i). The converse follows from B 3.5l and B 3.5n.

By A 4.2c, lemma B 3.5o is equivalent to the formula

$$H^{(q,\lambda)} = (\widetilde{P}^{(q,\lambda)}, P^{(q)}I^{(\mu-\lambda)}, I^{(\mu)} - P^{(q+1)}I^{(\mu-\lambda-1)}) \qquad \text{B 3.5p}$$

B 3.5q These results show that any complex of linear spaces corresponds uniquely to a set of proportional tensors of the type f_H and therefore also to f_G, f^G, and f^H, related by B 3.5c, B 3.5e, B 3.5f, B 3.5g.

B 3.5r Lemma. Let Λ be the complex of linear (r-1)-spaces determined by f_G, f^G, f_H, f^H and let $x^{\alpha} \in \mathcal{C}_{\alpha^*}$. Then conditions for the point x to be singular for Λ are given by $f_H\bullet(x)^{\lambda}I^{(\mu-\lambda)} = 0$ and $K^{(\mu-\lambda+1,\lambda)}\bullet f^H(x)^{\lambda} = 0$.

Proof. The condition for x to be singular for Λ is that the join of x to any(r-2)-space not containing it should belong to Λ. Let the (r-2)-space be given by the points $y^{\alpha}_{(1)}$, .., $y^{\alpha}_{(\lambda)}$. Then we have

CHAPTER 6

$f_H \bullet (x)^{\alpha} (y_{(2)})^{\alpha} .. (y_{(\lambda)})^{\alpha} = 0$ for all $y^{\alpha}_{(2)}, .., y^{\alpha}_{(\lambda)}$ if x is singular for Λ, whence $f_H \bullet (x)^{\alpha} (P^{(\alpha)})^{\lambda-1} = 0$. Thus $f_H \bullet (x)^{\alpha} I^{(n\alpha-\alpha)} = 0$ (by the symmetry of f_H).

If the (r-2)-space is given by the dual Grassmann tensor $q_{\alpha^{n-\lambda+1}}$ its join with x is given by $q \bullet I^{(n-\lambda)} x$ (see A 4.4 h). Thus, if x is singular for Λ, $(q \bullet I^{(n-\lambda)} x)^{\alpha} \bullet K^{(n-\lambda+\alpha)} \bullet f^H = 0$. This may be written $(q \bullet I^{(n-\lambda)} x)^{\alpha} \bullet M^{(n-\lambda,\alpha)} \bullet f^H = 0$, i.e. $(q)^{\alpha} \bullet M^{(n-\lambda+H,\alpha)} \bullet f^H (x)^{\alpha} = 0$, i.e. $(q)^{\alpha} \bullet K^{(n\lambda+\lambda+H,\alpha)} \bullet f^H (x)^{\alpha} = 0$ (see B 3.1m).
The result follows from B 3.1o and B 3.2c. B 3.6 Theorem.

The relations $q \bullet (I^{(k)} - N^{(k)}) = 0$ and $(q)^2 \bullet N^{(k+1)} I^{(k-1)} = 0$ form a basis for $\mathcal{U}_G(k)$.

Proof. Applying B 3.1g, B 3.1h, B 3.1i to B 3.5p, we get $G^{(\alpha,\lambda)} = (P^{(\alpha,\lambda)}, N^{(\alpha)} I^{(\alpha\lambda-\alpha)}, I^{(\alpha\lambda)} - N^{(\alpha\lambda+1)} I^{(\alpha\lambda-\alpha-1)})$. Hence the result follows from B 2.4c.

B 3.6a The result B 3.6 is the basis theorem for (dual) Grassmann coordinates since, as in A 4.4j, the relation $(q)^2 \bullet N^{(k+1)} I^{(k-1)} = 0$ is equivalent to the "quadratic p-relations." (See also Hodge and Pedoe (9) pp.315-321).

B 3.6b Second fundamental theorem of invariant theory. Theorem B 3.6 is also equivalent to the second main theorem of invariant theory, which states, essentially, that a basis for relations satisfied by the tensors ϵ and ϵ^* is given by $\epsilon^* \bullet (I - N^{(n)}) = 0$, $(\epsilon^*)^2 \bullet N^{(n+1)} I^{(n-1)} = 0$, $(I - N^{(n)}) \bullet \epsilon = 0$ and $N^{(n+1)} I^{(n-1)} (\epsilon)^2 = 0$, together with $\epsilon \epsilon^* = n! N^{(n)}$ (see Turnbull (21) pp.214,225). Since $\epsilon \epsilon^* = n! N^{(n)}$, any simple invariant need involve only ϵ or only ϵ^* (see B 1.3c).

B 3.7 **Theorem.** Let $\beta, \gamma, \delta, .., \zeta \in \mathfrak{G}$ and $a^{\gamma}_{\beta k}, b^{\delta}_{\beta k}, .., d^{\zeta}_{\beta m} \in \mathcal{C}$. Then a necessary and sufficient condition for the relation

$$a \bullet (x)^{k} = 0$$

to hold for all tensors x^{β} of $\mathcal{C}_{\beta *}$ for which

$$b \bullet (x)^{k} = 0, \ \ldots, \ d \bullet (x)^{m} = 0,$$

is that there exist integers $r, s \geqslant 0$ such that

$$(a)^{q} \, (bI^{(k_a - k)}_{(\beta)})^{t} \, .. \, (dI^{(k_a - m)}_{(\beta)})^{v} \bullet (P^{(k_a)}_{(\beta)})^{\lambda+1} \bullet N^{(k_a, \lambda+1)}_{(\beta)} = 0$$

for all $t, .., v \geqslant 0$ for which $t + .. + v = r$ and that for some such $t, .., v$

$$(bI^{(k_a - k)}_{(\beta)})^{t} \, .. \, (dI^{(k_a - m)}_{(\beta)})^{v} \bullet (P^{(k_a)}_{(\beta)})^{\lambda} \bullet N^{(k_a, \lambda)}_{(\beta)} \neq 0.$$

Proof. Let $z_{(\gamma, i)\gamma}, z_{(\delta, j)\delta}, \ldots$ where $i = 1, .., \nu(\gamma)$; $j = 1, .., \nu(\delta)$; .... be tensors spanning $\mathcal{C}_{\gamma}, \mathcal{C}_{\delta}, \ldots$ respectively. Then the condition $a \bullet (x)^{k} = 0$ is equivalent to

$$(z_{(\gamma,1)} \bullet a) \bullet (x)^{k} = 0, \ \ldots, \ (z_{(\delta, \nu(\gamma))} \bullet a) \bullet (x)^{k} = 0.$$

Similarly $b \bullet (x)^{k} = 0$ is equivalent to

$$(z_{(\delta,1)} \bullet b) \bullet (x)^{k} = 0, \ \ldots, \ (z_{(\delta, \nu(\delta))} \bullet b) \bullet (x)^{k} = 0,$$

and so on. Now the expressions $(z_{(\gamma,\delta} a) \bullet (x)^{k}$, $(z_{(\delta,j)} \bullet b) \bullet (x)^{k}$, ... are polynomials in the components of x in some frame $\mathcal{F}$. Thus, since $\mathfrak{A}$ is algebraically closed, the Hilbert zero theorem can be used to give the result that a necessary and sufficient condition for $a \bullet (x)^{k} = 0$ to hold for all x^{β} of $\mathcal{C}_{\beta *}$ for which $b \bullet (x)^{k} = 0, .., d \bullet (x)^{m} = 0$ is that there exist integers $s_i > 0$ such that for each $i = 1, .., \nu(\delta)$,

$$\left\{(z_{(\gamma,i)}\bullet a)\bullet(x)^{h}\right\}^{a_i} \equiv \left\{y_{(b,i,1)}\bullet(x)^{a_i-h}\right\}\left\{(z_{(\beta,1)}\bullet b)\bullet(x)^{h}\right\} + \ldots + \left\{y_{(b,i,\nu(\beta))}\bullet(x)^{a_i-h}\right\}\left\{(z_{(\beta,\nu(\beta))}\bullet b)\bullet(x)^{h}\right\}$$
$$+ \ldots + \left\{y_{(d,i,1)}\bullet(x)^{a_i-m}\right\}\left\{(z_{(\delta,1)}\bullet d)\bullet(x)^{m}\right\} + \ldots + \left\{y_{(d,i,\nu(\delta))}\bullet(x)^{a_i-m}\right\}\left\{(z_{(\delta,\nu(\delta))}\bullet d)\bullet(x)^{m}\right\}$$

is an identity in x^{β}, for some $y_{(b,i,1)},\ldots,y_{(d,i,\nu(\delta))} \in \mathcal{C}$. This condition can be written as

$$(z_{(\gamma,i)}\bullet a)^{a_i}\bullet(x)^{ha_i} \equiv (y_{(b,i)}\bullet bI_{(\beta)}^{(ha_i-h)})\bullet(x)^{ha_i} + \ldots + (y_{(d,i)}\bullet dI_{(\beta)}^{(ha_i-m)})\bullet(x)^{ha_i} \qquad (i = 1,\ldots,\nu(\gamma))$$

for some $y_{(b,1)}, \ldots, y_{(d,\nu(\delta))}$ of $\mathcal{C}$.

Putting $h = 1 + (s_1-1) + \ldots + (s_{\nu(\delta)}-1)$, this condition can be written as

$$(a)^{\wedge}\bullet(x)^{ha} \equiv (y_{(b)}\bullet bI_{(\beta)}^{(ha-h)})\bullet(x)^{ha} + \ldots + (y_{(d)}\bullet dI_{(\beta)}^{(ha-m)})\bullet(x)^{ha}$$

for some $y_{(b)},\ldots,y_{(d)} \in \mathcal{C}$. This follows from A 3.5e because the products $z_{(\gamma,p_1)}\ldots z_{(\delta,p_A)}$ with $p_1,\ldots,p_A$ taken from $1,\ldots,\nu(\gamma)$, span the space $\mathcal{C}_{\gamma A}$ and each such product contains at least s_i of the $z_{(\gamma,i)}$'s, for some i. (It is really the <u>extension</u> of the Hilbert zero theorem that is being used here, see van der Waerden (23) p.6).

Hence, by B 1.2, a necessary and sufficient condition for $a\bullet(x)^{h} = 0$ to hold for all x^{β} of $\mathcal{C}_{\beta}$* for which $b\bullet(x)^{h} = 0,\ldots,d\bullet(x)^{m} = 0$ is the existence of an integer $s > 0$ and elements $y_{(b)}, \ldots, y_{(d)}$ of $\mathcal{C}$ such that

$$(a)^{s}\bullet P_{(\beta)}^{(ha)} = y_{(b)}\bullet bI_{(\beta)}^{(ha-h)}\bullet P_{(\beta)}^{(ha)} + \ldots + y_{(d)}\bullet dI_{(\beta)}^{(ha-m)}\bullet P_{(\beta)}^{(ha)}.$$

Thus, the result follows from A 3.5d.

B 3.7a <u>Corollary</u>. Let $a_{\beta A}^{\gamma}$, $b_{\beta A}^{\delta} \in \mathcal{C}$. Then a necessary and sufficient condition for the relation $a\bullet(x)^{h} = 0$ to hold for all tensors x^{β} of $\mathcal{C}_{\beta}$* for which $b\bullet(x)^{h} = 0$ is the existence of integers $r, s \geqslant 0$ such that

$$(a)^{s}\ (bI_{(\beta)}^{(ha-h)})^{\wedge}\bullet(P_{(\beta)}^{(ha)})^{\wedge r}\bullet N_{(\beta)}^{(ha,\wedge r+1)} = 0$$

and

$$(bI_{(\beta)}^{(ha-h)})^{\wedge}\bullet(P_{(\beta)}^{(ha)})^{\wedge}\bullet N_{(\beta)}^{(ha,\wedge)} \neq 0.$$

B 3.7b <u>Corollary</u>. If $b^{\delta}_{\beta k},..,d^{\xi}_{\beta m} \in \mathcal{C}$, then a necessary and sufficient condition for the existence of a non-zero element x^{β} of $\mathcal{C}_{\beta *}$ which satisfies all of $b_{\bullet}(x)^{k} = 0,..,\ d_{\bullet}(x)^{m} = 0$, is that for each integer $s \geqslant \max(k,..,m)$,

$$(bI^{(a-k)}_{(\beta)})^{t}..(dI^{(a-m)}_{(\beta)})^{v}\bullet(P^{(a)}_{(\beta)})^{\lambda}\ N^{(a,\lambda)}_{(\beta)} = 0$$

for all $t,..,v \geqslant 0$ for which $t + .. + v = r = \binom{\Delta+\nu(\beta)}{\Delta}$.

<u>Proof</u>. Put $a = I_{(\rho)}$ in B 3.7. Then B 3.7 tells us that a necessary and sufficient condition for all common solutions x^{β}, of $b_{\bullet}(x)^{k} = 0,..,d_{\bullet}(x)^{m} = 0$ also to satisfy $x = 0$, is that there exist integers $r,\ s \geqslant 0$ such that for each $t,..,v \geqslant 0$ for which $t + .. + v = r$

$$I^{(a)}_{(\beta)}(bI^{(a-k)}_{(\beta)})^{t}..(dI^{(a-m)}_{(\beta)})^{v}\bullet(P^{(a)}_{(\beta)})^{\lambda+1}\bullet N^{(a,\lambda+1)} = 0$$

but for some $t,..,\ v$

$$(bI^{(a-k)}_{(\beta)})^{t}..(dI^{(a-m)}_{(\beta)})^{v}\bullet(P^{(a)}_{(\beta)})^{\lambda}\bullet\ N^{(a,\lambda)}_{(\beta)} \neq 0.$$

Now, by A 1.5u, $\langle P^{(a)}_{(\beta)}\rangle = \binom{\Delta+\nu(\beta)}{\Delta}$. Hence, putting $e = P^{(a)}_{(\beta)}$ and $t = r-1$ in A 4.4m we see that unless $r = \binom{\Delta+\nu(\beta)}{\Delta}$, the first of the above two expressions cannot vanish unless the second does also. Hence $r = \binom{\Delta+\nu(\beta)}{\Delta}$ and with this value for r we have $(P^{(\lambda)}_{(\beta)})^{\lambda+1}\bullet N^{(\lambda+1)}_{(\beta)} = 0$ in any case, since $P^{(a)}_{(\beta)}$ has rank r.

Thus, it follows that a necessary and sufficient condition for $b_{\bullet}(x)^{k} = 0,..,\ d_{\bullet}(x)^{m} = 0$ to have a non-trivial common solution is that given in B 3.7b.

B 3.7c Corollary B 3.7b shows that the sets of tensors $(b^{\delta}_{\beta k},..,d^{\xi}_{\beta m})$ for which $b_{\bullet}(x)^{k} = 0,..d_{\bullet}(x)^{m} = 0$ have a non-trivial solution, constitute a geometric variety in the space $\mathcal{C}_{\beta k\delta^{*}} \times .. \times \mathcal{C}_{\beta m\zeta^{*}}$. This is the <u>resultant</u> variety of b,..,d.

If $\delta = .. = \zeta = 1$, the expressions given in B 3.7b, when expressed in terms of components, are resultant (inertia) forms of the forms $b_\bullet(x)^{k}$, .., $d_\bullet(x)^{m}$.

(The idea of obtaining the resultant system from the Hilbert zero theorem, is due to van der Waerden (24)).

B 3.7d The resultant of n forms in n variables. It is well known that n homogeneous polynomials in n variables have a unique resultant form (that is, unique except for a constant of proportionality), with the property that it is a basis for the resultant system of the given polynomials, its vanishing being the necessary and sufficient condition for the polynomials to have a non-trivial common zero. (see van der Waerden (23).)

Let $b_{(1)\alpha_1 \ldots \alpha_{k_1}}, \ldots, b_{(n)\alpha_1 \ldots \alpha_{k_n}}$ be n symmetric elements of $\mathcal{C}$ (that is, subject to $b_{(1)} = b_{(1)} \bullet P^{(k_1)}, \ldots, b_{(n)} = b_{(n)} \bullet P^{(k_n)}$). Then $b_{(1)\bullet}(x)^{k_1}, \ldots, b_{(n)\bullet}(x)^{k_n}$ are n forms in the components of the tensor x^α of $\mathcal{C}$, in a frame $\mathcal{F}$ of $\mathcal{C}$. These components are $\nu(\alpha) = n$ in number. Hence there is a resultant form

$$\phi = \phi(b_{(1)}, \ldots, b_{(n)})$$

in the components of $b_{(1)}, \ldots, b_{(n)}$ in $\mathcal{F}$, whose vanishing is a necessary and sufficient condition for the existence of a non-zero x^α of $\mathcal{C}$ satisfying all of $b_{(1)\bullet}(x)^{k_1} = 0, \ldots, b_{(n)\bullet}(x)^{k_n} = 0$. Now, this condition is obviously independent of the choice of $\mathcal{F}$. Therefore, by B 1.4, we can take ϕ to be of the form

$$\phi(b_{(1)}, \ldots, b_{(n)}) = \{(b_{(1)})^{k_1} \ldots (b_{(n)})^{k_n}\}_\bullet R^{(k_1, \ldots, k_n)}{}_\bullet(\epsilon)^{k_1 \ldots k_n}$$

where $R^{(k_1,\ldots,k_m)} \in \mathfrak{R}$ and $h_1,\ldots,h_m$ are such that $h_1k_1 = \ldots = h_mk_m = k_1k_2\ldots k_m$, since the degrees of ϕ in the coefficients are as given. Also the weight of ϕ must be $k_1\ldots k_m$, as given, in order to ensure that the expression is a scalar.

By B 1.2a and B 3.2c (see B 3.2f), we can suppose that $R^{(k_1,\ldots,k_m)}$ has the symmetry given by

$$R^{(k_1,\ldots,k_m)} = \{P^{(k_1,k_1)}\bullet(P^{(k_1,1)})^{k_1}\}\ldots\{P^{(k_m,k_m)}\bullet(P^{(k_m,1)})^{k_m}\}\bullet R^{(k_1,\ldots,k_m)}\bullet G^{(m,k_1\ldots k_m)}$$

and this will define $R^{(k_1,\ldots,k_m)}_{(\alpha)}$ uniquely, except for proportionality, as an element of $\mathfrak{A}$. It will also define $R^{(k_1,\ldots,k_m)}$ uniquely as an element of $\mathfrak{F}$ since $G^{(m,k_1\ldots k_m)}\bullet Q^{(mk_1\ldots k_m)}_m = 0$.

Since $R^{(k_1,\ldots,k_m)}$ defines a unique element of $\mathfrak{F}$, it will also be possible to define corresponding elements $R^{(k_1,\ldots,k_m)}$ for each m.

B 3.7e Remark. Let $b_{(1)\alpha^{k_1}},\ldots,b_{(m)\alpha^{k_m}} \in \mathfrak{C}$. Then according to B 3.7b, the necessary and sufficient condition for $b_{(1)}\bullet(x)^{k_1} = 0,\ldots, b_{(m)}\bullet(x)^{k_m} = 0$ to have a common non-trivial solution x^α is that for all suitable $s, t_1,\ldots,t_m$ with $t_1 + \ldots + t_m = r = \binom{a+m}{a}$,

$$(b_{(1)}I^{(a-k_1)})^{t_1}\ldots(b_{(m)}I^{(a-k_m)})^{t_m}\bullet(P^{(a)})^{\wedge}\bullet N^{(a,\lambda)} = 0.$$

Now it can be shown that these relations for which

$$s = 1 + (k_1 - 1) + \ldots + (k_m - 1),$$

are sufficient to ensure such a solution. (See van der Waerden (23) p.12.)

Also, by A 4.4n, $(P^{(a)})^{\wedge}\bullet N^{(a,\lambda)} = \lambda\epsilon^{(a,\lambda)}\epsilon^{(a,\lambda)*}$ $(\lambda \neq 0)$ where $\epsilon^{(a,\lambda)}$ is a certain α^Δ-Grassmann tensor. It is skew-symmetric in its r groups of s symmetric indices and can be constructed out of ϵ tensors.

CHAPTER 6

Thus, the condition on the b's can be written as

$$(b_{(1)} I^{(a-k_1)})^{t_1} \ldots (b_{(m)} I^{(a-k_m)})^{t_m} \bullet \epsilon^{(a,n)} = 0$$

for each $t_1, \ldots, t_m \geqslant 0$ such that $t_1 + \ldots + t_m = r = \binom{a+m}{a}$ and $s = k_1 + \ldots + k_m - n + 1$.

Now since the resultant $\phi(b_{(1)}, \ldots, b_{(m)})$ is a factor of any other inertia form of $b_{(1)} \bullet (x)^{k_1}, \ldots, b_{(m)} \bullet (x)^{k_m}$, it follows that

$$(b_{(1)} I^{(a-k_1)})^{t_1} \ldots (b_{(m)} I^{(a-k_m)})^{t_m} \bullet \epsilon^{(a,n)} \equiv \phi(b_{(1)}, \ldots, b_{(m)}) \{(b_{(1)})^{t_1-k_1} \ldots (b_{(m)})^{t_m-k_m} I^{(na-k_1t_1-\ldots-k_mt_m)} \bullet w\}$$

for each $t_1, \ldots, t_m$, where w is an element of $\mathcal{C}$ (independent of $b_{(1)}, \ldots, b_{(m)}$ but depending on $t_1, \ldots, t_m$). (If for some i, $t_i < h_i = \frac{k_1 \ldots k_m}{k_i}$, then l.h.s. $= 0$.) It is therefore reasonable to suppose that an expression for ϕ might be found which bears some similarity to the left-hand side of this equation, and indeed this appears to be the case (at least in the simpler cases).

If $n = 2$, there is the following expression for ϕ :

$$\phi(b_{(1)}, b_{(2)}) = \epsilon^{(k_2-1, k_1)*} \epsilon^{(k_1-1, k_2)*} \bullet (b_{(1)} I^{(k_2-1)})^{k_1} (b_{(2)} I^{(k_1-1)})^{k_2} \bullet \epsilon^{(k_1+k_2-1, k_1+k_2)}$$

It is possible to relate this expression directly to the Sylvester determinant.

The construction of ϕ is not so simple if $n > 2$, but if $n = 3$, there is the following expression for the resultant of three ternary quadratics $a \bullet (x)^2$, $b \bullet (x)^2$, $c \bullet (x)^2$:

$$\phi(a,b,c) = a_1 a_3 a_5 a_7 b_9 b_{11} b_{13} b_{15} c_{17} c_{19} c_{21} c_{23} \, \epsilon^{(2,6)*}_{2,4,6,8,25,27} \, \epsilon^{(2,6)*}_{10,12,14,16,26,29} \, \epsilon^{(2,6)*}_{18,20,22,24,28,30} \, \epsilon^{(6,15)\,1,2,3,\ldots,30},$$

where the numerical indices $1, 2, \ldots, 30$ stand for $\bar{\gamma}_1, \bar{\gamma}_2, \ldots, \bar{\gamma}_{30}$ respectively, with $\gamma = \alpha^2$, and contraction takes place in each case.

The tensor $\epsilon^{(2,6)*}$ has components $\epsilon^{(2,6)*}_{i_1 i_2, \ldots, i_{11} i_{12}} = \pm 1$ or 0 according as the unordered pairs $(i_1,i_2),\ldots,(i_{11},i_{12})$ constitute an even, odd permutation of $(1,1),(1,2),(1,3),(2,2),(2,3),(3,3)$ or otherwise. The components of $\epsilon^{(4,15)}$ are similarly defined.

B 4 Cayley Forms.

In this chapter I shall give some applications of the precedding theory to the study of Cayley forms. In particular, it will be possible to give a partial solution to the problem of finding the Cayley form of the intersection of two varieties in terms of their Cayley forms. Also, an expression for the relations which are necessary and sufficient for a form to be a Cayley form will be given.

B 4.1 Cayley tensors. Let $|\mathcal{K}|$ be an unmixed algebraic variety in the projective (n-1)-space $\mathfrak{P}$, $\mathcal{K}$ being the set of elements x^α of $\mathcal{C}_{\alpha *}$ which correspond to the points of $|\mathcal{K}|$. Now, since $|\mathcal{K}|$ is unmixed, it has a Cayley form $\Psi = \Psi(u_{(1)},\ldots,u_{(n-k)})$ which is a polynomial homogeneous in the components of each of the elements $u_{(1)\alpha},\ldots,u_{(n-k)\alpha}$ of $\mathcal{C}$. These elements represent primes in $\mathfrak{P}$ and the relation $\Psi = 0$ is a necessary and sufficient condition for the intersection of these primes to meet $|\mathcal{K}|$. The dimension of $|\mathcal{K}|$ is n-k-1 and its order is the degree r of the polynomial Ψ (see Hodge and Pedoe (10) pp. 32-41).

The linear space of intersection of $u_{(1)},\ldots,u_{(n-k)}$ can also be

CHAPTER 6

represented by its dual Grassmann tensor $q = u_{(1)}..u_{(m-k)} \cdot N^{(m-k)}$, by the corresponding Grassmann tensor $p = \frac{1}{k!} qI^{(k)} \cdot \epsilon$, or by a set of vectors $x^{\alpha}_{(1)}, .., x^{\alpha}_{(k)}$ where $p = N^{(k)} \cdot x_{(1)}..x_{(k)}$. The condition $\Psi = 0$ states that this linear space belongs to the complex of (k-1)-spaces which meet $|\mathcal{K}|$. Thus, by B 3.5q, we have found tensors f_G, f^G, f_H and f^H, related by B 3.5c, B 3.5e, B 3.5f and B 3.5g, which are uniquely determined by this complex and therefore by $|\mathcal{K}|$, except for a scalar multiplier, and such that

$$\Psi \equiv (u_{(1)})^{\wedge}..(u_{(m-k)})^{\wedge} \cdot f^H \equiv (q)^{\wedge} \cdot f^G \equiv f_G \cdot (p)^{\wedge} \equiv f_H \cdot (x_{(1)})^{\wedge}..(x_{(k)})^{\wedge}.$$

I shall call f_G, f^G, f_H and f^H Cayley tensors corresponding to $|\mathcal{K}|$.

B 4.1a If, in particular, $|\mathcal{K}|$ is a linear space, then $f_G = f^i_H$ and $f^G = f^H$ are respectively a dual α-Grassmann tensor and α-Grassmann tensor corresponding to $|\mathcal{K}|$. It turns out, also, that f_G and $(-1)^{k(m-k)} f^G$ are related as are p and q in A 4.5d. As with linear spaces I shall frequently use the tensor symbol f, to denote the variety $|\mathcal{K}|$. However, the symbol f will be taken to mean the multiplicative variety associated with its Cayley tensor f_G, rather than the point set variety $|\mathcal{K}|$.

B 4.1b Let p′ and p″ be the Grassmann tensors given by $p' = N^{(l)} \cdot x_{(1)}..x_{(l)}$ and $p = N^{(k-l)} \cdot x_{(l+1)}..x_{(k)}$. Then the linear space p is the join of the disjoint spaces p′ and p″. Also $p = N^{(k)} \cdot p'p''$. Now $f_G = f_G \cdot G^{(k,k)} = f_G \cdot (N^{(k)})^{\wedge}$. Hence $f_G \cdot (p'p'')^{\wedge} = f_G \cdot (p)^{\wedge}$. Suppose that p′ does not meet f (= $|\mathcal{K}|$). Then a necessary and sufficient condition for p″ to meet the cone projecting f from the space p′ is $f_G \cdot (p'p'')^{\wedge} = 0$. Hence, a Cayley tensor corresponding to

this cone is given by $f_G \bullet (p' I^{(k-l)})^\wedge \bullet G^{(k-l,\lambda)}$. But this may be written $f_G \bullet G^{(k,\lambda)} \bullet D^{(p:k;\lambda)} *_\bullet (p')^\wedge G^{(k-l,\lambda)}$ by A 4.5f and B 3.5a. Thus, by B 3.2c and the conjugate of B 3.5d ($G^{(\cdot,\cdot)*} = G^{(\cdot,\cdot)}$), the final $G^{(k-l,\lambda)}$ is redundant, so that this Cayley tensor may be written $f_G \bullet (p' I^{(k-l)})^\wedge$. Similarly, $f_H \bullet (x_{(1)}) .. (x_{(l)}) I^{(k_\lambda - l_\lambda)}$ is also a Cayley tensor corresponding to this cone.

B 4.1c If the space p' meets the variety f, then this cone is the whole space and we have $f_G \bullet (p' I^{(k-l)})^\wedge = 0$ and $f_H \bullet (x_{(1)}) .. (x_{(l)}) I^{(k_\lambda - l_\lambda)} = 0$. It follows that necessary and sufficient conditions for the linear space p', which is the join of the points $x_{(1)}, .., x_{(l)}$, to meet the variety corresponding to f_G and f_H are

$$f_G \bullet (p' I^{(k-l)})^\wedge = 0 \text{ and } f_H \bullet (x_{(1)})^\wedge .. (x_{(l)})^\wedge I^{(k_\lambda - l_\lambda)} = 0.$$

The particular case of this for which $l = k$ gives the condition already considered, for the $(k-1)$-space p to meet f. If $l = 1$, we get

B 4.1d **Lemma.** $x^\alpha \in \mathcal{K}$ if and only if $f_H \bullet (x)^\wedge I^{(k_\lambda - \lambda)} = 0$.

B 4.1e **Lemma.** If $g_{\alpha^{k_\lambda}} \in \mathcal{C}$ where $0 \neq g = g \bullet H^{(\lambda, k_\lambda)}$ and $g \bullet (x)^\wedge I^{(k_\lambda - \lambda)} = 0$ for all $x^\alpha \in \mathcal{K}$, then g is proportional to f_H.

Proof. By B 3.5i, g determines a complex Λ of $(k-1)$-spaces of degree r. By B 3.5r each point of $|\mathcal{K}|$ is singular for Λ. Hence Λ is the complex of $(k-1)$-spaces meeting $\mathcal{K}$. The result follows. There is a corresponding result for f_G (see B4.1c). B 4.1f Let $q' = u_{(1)} .. u_{(m)} \bullet N^{(m)}$ be a dual Grassmann tensor ($m < n-k$). Then, by arguments similar to that of B 4.1b, it follows that Cayley tensors corresponding to the intersection of q' with f are given by

$(q'I^{(n-k-m)})^{\wedge}\cdot f^{G}$, $(u_{(1)})^{\wedge}..(u_{(m)})^{\wedge}I^{(n(n-k-m))}\cdot f^{H}$ and $f_{G}\cdot\{q'I^{(k)}\cdot N^{(m+k)}\}$, provided that this intersection has dimension n-m-k-1.

B 4.1g Remark. In accordance with B 2.4h, the homogeneous ideal of polynomials $a\cdot(x)^{A}$ which vanish for all x^{α} of $\mathcal{K}$, corresponds uniquely to a sequence of projectors $e^{(0)}$, $e^{(1)}$, $e^{(2)}$, ... satisfying $(e^{(t)}I^{(A-t)})\cdot e^{(A)} = e^{(A)} = P^{(A)}\cdot e^{(A)}$ for which $a\cdot e^{(A)} = 0$ holds if and only if $a\cdot(x)^{A}$ belongs to the ideal. Now any algebraic variety corresponds uniquely to such a sequence of projectors. Hence it should be possible to obtain its Cayley tensors from them.

Now it follows from B 4.1d that $f_{H}\cdot e^{(\lambda)}I^{(k\lambda-\lambda)} = 0$ so that $f_{H}\cdot\{I^{(\lambda)} - e^{(\lambda)}\}I^{(k\lambda-\lambda)} = f_{H}$. Hence by A 4.2c, $f_{H} = f_{H}\cdot(I^{(k\lambda)} - e^{(\lambda)}I^{(k\lambda-\lambda)}, H^{(\lambda,k)}) \neq 0$. Now let $g_{\alpha k\lambda}$ be any element of $\mathcal{C}_{\alpha k\lambda}$ such that $0 \neq g\cdot(I^{(k\lambda)} - e^{(\lambda)}I^{(k\lambda-\lambda)}, H^{(\lambda,k)}) = g'$, say. Then we have $g'\cdot H^{(\lambda,k)} = g'$ and $g'\cdot e^{(\lambda)}I^{(k\lambda-\lambda)} = 0$. But $(x)^{\lambda} = e^{(\lambda)}\cdot(x)^{\lambda}$ for all $x \in \mathcal{K}$ (see B 2.3a). Thus $g'\cdot(x)^{\lambda}I^{(k\lambda-\lambda)} = 0$ for all $x \in \mathcal{K}$. Hence, by B 4.1e, g' is proportional to f_{H}.

This shows that a Cayley tensor g_{H} for $|\mathcal{K}|$ can be obtained from $e^{(\lambda)}$ and an arbitrary element g of $\mathcal{C}_{\alpha k\lambda}$ by putting

$$g_{H} = g\cdot(I^{(k\lambda)} - e^{(\lambda)}I^{(k\lambda-\lambda)}, H^{(\lambda,k)})$$

$(= g\cdot\{H^{(\lambda,k)} - H^{(\lambda,k)}\cdot(H^{(\lambda,k)}\cdot e^{(\lambda)}I^{(k\lambda-\lambda)})^{\dagger}\}$ see A 4.3c.)

B 4.2 The van der Waerden varieties.

Let $X = X(u_{(1)}, .., u_{(n-k)})$ be a polynomial which is homogeneous of degree r in the components, in some frame $\mathcal{F}$, of $u_{(1)\alpha}, .., u_{(n-k)\alpha}$. Now, there is a

set of relations connecting the coefficients of X which represents a necessary and sufficient condition that X be the Cayley form of some multiplicative algebraic variety in $\mathfrak{B}$. These relations take the form of the vanishing of a set of polynomials which are homogeneous in the coefficients of X (see (10) pp.51-57).

These polynomials have rational coefficients, since no mention of the ground field is made in obtaining them and they exist for any ground field without characteristic, e.g. for the rationals. This can also be seen by examining the relations themselves, as given in B 4.8 .

Now let $X(u_{(1)},..,u_{(n-k)}) \equiv (u_{(1)})^{\wedge}..(u_{(n-k)})^{\wedge} \bullet a$ where $a = (P^{(\wedge)})^{n-k} \bullet a$. Then the condition for X to be a Cayley form is given by the vanishing of a set of polynomials over the rationals, homogeneous in the components of a in the frame $\mathcal{F}$. However, this condition for a to be a Cayley tensor is clearly independent of $\mathcal{F}$. Hence, by B 1.3, the condition can be written as

$$F \bullet (a)^{\Delta} = 0$$

where F is some element of $\mathcal{R}_{(\alpha\alpha^{*})^{n(n-k)}}$.

Thus, the condition for $a \neq 0$ to be a Cayley tensor of the type a^{H} corresponding to a variety of order r and dimension n-k-1 is that it belong to a certain geometric variety in $\mathcal{C}_{\alpha^{*n(n-k)}}$.

If a^{G} is related to $a = a^{H}$ by $a^{G} = K^{(n-k,n)\dagger} \bullet a^{H}$ as in B 3.5g, then a^{G} is

an appropriate Cayley tensor whenever $a^H \neq 0$ satisfies $F\bullet(a^H)^A = 0$. Conversely if a^G is a Cayley tensor and $a^H = K^{(n-k-1)H}\bullet a^G$ as in B 3.5f, then a^H satisfies $F\bullet(a^H)^A = 0$. Thus,it follows that the condition for $a^G \neq 0$ to be an appropriate Cayley tensor corresponding to a variety of order r and dimension n-k-1 is that it belong to the geometric variety given by

$$F\bullet(K^{(n-k-1)H})^A\bullet a^G = 0,\ a^G = G^{(n-k-1)}\bullet a^G$$

Similarly, it can be shown that the Cayley tensors of type a_G and a_H which correspond to some algebraic variety of order r and dimension n-k-1, also constitute geometric varieties (in $\mathcal{C}_{\alpha k_1}$).

B 4.2a Notation. I shall denote by $\mathcal{W}_H(k,r)$, $\mathcal{W}_G(k,r)$, $\mathcal{W}^H(k,r)$ and $\mathcal{W}^G(k,r)$, the respective geometric varieties of the Cayley tensors f_H, f_G, f^H and f^G which correspond to some unmixed multiplicative variety of order r and dimension n-k-1.

The varieties $|\mathcal{W}_H(k,r)|$, .., $|\mathcal{W}^G(k,r)|$ are all projectively equivalent, because of the linear nature of B 3.5c, B 3.5e, B 3.5f and B 3.5g (see (10) p.57.) They are reducible if r, k, n-k > 1.

I shall use the following notation for the projectors associated with the $\mathcal{W}$'s :

$$E^{(a)}_{\mathcal{W}_H(k,r)} = W_H^{(a,k;a)},\ E^{(a)}_{\mathcal{W}_G(k,r)} = W_G^{(k,r;a)},\ E^{(a)}_{\mathcal{W}^H(k,r)} = W^{H(r,n-k;a)},\ E^{(a)}_{\mathcal{W}^G(k,r)} = W^{G(n-k,r;a)}. \qquad \text{B 4.2b}$$

The bracketed indices of the W's indicate the grouping of their tensor indices.

The W's are all elements of $\mathcal{R}$, by B 2.2r. By B 2.3a and B 2.3 we have the following two results.

B 4.2c Lemma. Let $f_H \in \mathcal{W}_H(k,r)$, $f_G \in \mathcal{W}_G(k,r)$, $f^H \in \mathcal{W}^H(k,r)$, $f^G \in \mathcal{W}^G(k,r)$. Then, for each $s \geqslant 0$,

$(f_H)^s \cdot W_H^{(\Lambda,\bar{\Lambda};s)} = (f_H)^s$, $(f_G)^s \cdot W_G^{(\bar{\Lambda},\Lambda;s)} = (f_G)^s$, $W^{H(\Lambda,n-\bar{\Lambda};s)} \cdot (f^H)^s = (f^H)^s$, $W^{G(n-\bar{\Lambda},\Lambda;s)} \cdot (f^G)^s = (f^G)^s$.

B 4.2d Lemma. Let $a_{\beta}^{\alpha_{\sim k}}$, $b_{\alpha_{\sim(n-k)}}^{\beta} \in \mathcal{C}$. Then the equations

$$W_H^{(\Lambda,\bar{\Lambda};s)} \cdot a = 0,\quad W_G^{(\bar{\Lambda},\Lambda;s)} \cdot a = 0,\quad b \cdot W^{H(\Lambda,n-\bar{\Lambda};s)} = 0,\quad b \cdot W^{G(n-\bar{\Lambda},\Lambda;s)} = 0$$

are the respective necessary and sufficient conditions for

$$(f_H)^s \cdot a = 0,\quad (f_G)^s \cdot a = 0,\quad b \cdot (f^H)^s = 0,\quad b \cdot (f^G)^s = 0$$

for all $f_H \in \mathcal{W}_H(k,r)$, $f_G \in \mathcal{W}_G(k,r)$, $f^H \in \mathcal{W}^H(k,r)$, $f^G \in \mathcal{W}^G(k,r)$ respectively.

B 4.3 Reducible varieties

Let f_G and f^G be Cayley tensors of a multiplicative variety f of order r and dimension n-k-1. Suppose that f is reducible into the two component varieties g and h of respective orders s and t (where s + t = r). Now, the Cayley form of f is the product of those of g and h, so that we have the identities

$$f_G \cdot (p)^r \equiv \{g_G \cdot (p)^s\}\{h_G \cdot (p)^t\} \quad \text{and} \quad (q)^r \cdot f^G \equiv \{(q)^s \cdot g^G\}\{(q)^t \cdot h^G\}$$

in the Grassmann tensor p and the dual Grassmann tensor q, where g_G, g^G, h_G and h^G are Cayley tensors of g and h.

Hence, by B 3.2c,

$$f_G = g_G h_G \cdot G^{(k,\Lambda)} \text{ and } f^G = G^{(n-k,\Lambda)} \cdot g^G h^G.$$

Similarly, if f is the union of the component varieties $f_{(1)}, \ldots f_{(l)}$, of equal dimensions, then

$$f_G = \{f_{(1)G} \ldots f_{(l)G}\} \cdot G^{(k,\Lambda)} \text{ and } f^G = G^{(n-k,\Lambda)} \cdot \{f_{(1)}^G \ldots f_{(l)}^G\}.$$

Thus, if in particular f consists of linear spaces, then f_G and f^G are respectively of the form

$$f_G = q_{(1)} \ldots q_{(\Lambda)} \cdot G^{(k,\Lambda)} \text{ and } f^G = G^{(n-k,\Lambda)} \cdot p_{(1)} \ldots p_{(\Lambda)}$$

where $q_{(1)}, \ldots, q_{(\Lambda)}$ are dual Grassmann tensors and $p_{(1)}, \ldots, p_{(\Lambda)}$ are corresponding Grassmann tensors (see B 4.1a).

Now the tensors f_G of the above form constitute a geometric subvariety of $\mathcal{W}_G(k,r)$. To see this, observe that the complex of linear spaces which meet an unmixed variety is also the dual of such a complex, if and only if the variety is reducible into linear spaces. Hence, the condition that f_G should be a Cayley tensor of such a variety is that $f_G \in \mathcal{W}_G(k,r)$ and that $f_G \in \{\mathcal{W}^G(n-k,r)\}^*$, that is, that f_G belong to the intersection of two geometric varieties.

B 4.3a <u>Notation</u>. I shall use the notation $\mathcal{U}_H(k,r), \mathcal{U}_G(k,r), \mathcal{U}^H(k,r), \mathcal{U}^G(k,r)$ for those geometric subvarieties of $\mathcal{W}_H(k,r), \mathcal{W}_G(k,r), \mathcal{W}^H(k,r), \mathcal{W}^G(k,r)$ respectively, which correspond to algebraic varieties in $\mathfrak{B}$ which consist of sets of linear spaces.

Then

$$\mathcal{U}_G(k,r) = \{\mathcal{U}^G(n-k,r)\}^* = \mathcal{W}_G(k,r) \frown \{\mathcal{W}^G(n-k,r)\}^*$$

and similarly

$$\mathcal{U}_H(k,r) = \{\mathcal{U}^H(n-k,r)\}^* = \mathcal{W}_H(k,r) \frown \{\mathcal{W}^H(n-k,r)\}^*. \qquad \text{B 4.3b}$$

In particular, if $k = n-1$, we have

$$\mathcal{W}_G(n-1,r) = \mathcal{U}_G(n-,r) \text{ and } \mathcal{W}_H(n-1,r) = \mathcal{U}_H(n-1,r)$$

and also

$$\mathcal{W}^G(n-1,r) = \mathcal{U}^G(n-1,r) = \mathcal{W}^H(n-1,r) = \mathcal{U}^G(n-1,r) \qquad \text{B 4.3c}$$

since any zero-dimensional variety is a set of points and since $K^{(1,\lambda)} = p^{(\lambda)}$.

If $r = 1$, the $|\mathcal{U}|$'s and $|\mathcal{W}|$'s are Grassmann varieties and we have

$$\mathcal{W}_G(k,1) = \mathcal{W}_H(k,1) = \mathcal{U}_G(k,1) = \mathcal{U}_H(k,1) = \mathcal{U}_G(k)$$

and

$$\mathcal{W}^G(k,1) = \mathcal{W}^H(k,1) = \mathcal{U}^G(k,1) = \mathcal{U}^H(k,1) = \mathcal{U}^G(k). \qquad \text{B 4.3d}$$

If $r \neq 1$ and $n-k \neq 1 \neq k$, then $|\mathcal{U}_G(k,r)|$ is in fact a proper irreducible component of $|\mathcal{W}_G(k,r)|$.

Putting

$$E^{(a)}_{\mathcal{U}_G(k,\lambda)} = U^{(k,\lambda,a)}$$

we also have

$$E^{(a)}_{\mathcal{U}^G(n-k,\lambda)} = U^{(k,\lambda,a)} \qquad \text{B 4.3e}$$

for each s, k, r, $n-k \geqslant 1$, by B 4.3b.

B 4.3f Lemma. For each s, k, r, $n-k \geqslant 1$, the tensor $U^{(k,\lambda,a)}$ is the unique projector having the property that for each $a^{\alpha_{\lambda}^{\ldots k}} \in \mathcal{C}$, the equation

$$U^{(k,\lambda,a)} \cdot a = 0$$

CHAPTER 6

is a necessary and sufficient condition for

$$\{q_{(1)} \ldots q_{(\lambda)} \cdot G^{(k,\lambda)}\}^{\alpha} \cdot a = 0$$

for each set $q_{(1)}, \ldots, q_{(\lambda)}$ of dual α-Grassmann tensors in $\mathcal{C}_{\alpha^k}$.

It is also the unique projector having the corresponding property for α-Grassmann tensors.

Proof. This is just B 4.3e applied to B 2.3.

B 4.3g Notation. For each k, r, s $\geqslant$ 0, I define

$$S^{(k,\alpha,\lambda)} = \{(G^{(k,\lambda)})^{\alpha} \cdot M^{(k;\lambda,\alpha)} \cdot (G^{(k,\alpha)})^{\lambda}\}^{\dagger},$$

so that $S^{(k,\alpha,\lambda)}_{(\alpha)} \in \mathfrak{P}_{\alpha}$. Put

$$S^{(\alpha,\lambda)} = S^{(1;\alpha,\lambda)} = \{(P^{(\lambda)})^{\alpha} \cdot M^{(\lambda,\alpha)} \cdot (P^{(\alpha)})^{\lambda}\}^{\dagger}.$$

B 4.3h Lemma. For each s, k, r, n-k $\geqslant$ 1

$$U^{(k,\lambda,\alpha)} = S^{(k,\lambda,\alpha)} \cdot \{(G^{(k,\alpha)})^{\lambda} \cdot M^{(k;\alpha,\lambda)} \cdot (G^{(k,\lambda)})^{\alpha}\} = \{(G^{(k,\lambda)})^{\alpha} \cdot M^{(k;\lambda,\alpha)} \cdot (G^{(k,\alpha)})^{\lambda}\} \cdot S^{(k,\alpha,\lambda)}.$$

Proof. Let $a^{\alpha\ldots\alpha}_{\beta\ldots\beta} \in \mathcal{C}$ and let $q_{(1)\alpha_1\ldots\alpha_k}, \ldots, q_{(\lambda)\alpha_1\ldots\alpha_k}$ be dual α-Grassmann tensors. Now, by B 3.1d

$$(q_{(1)} \ldots q_{(\lambda)} \cdot G^{(k,\lambda)})^{\alpha} \cdot a = (q_{(1)})^{\alpha} \ldots (q_{(\lambda)})^{\alpha} \cdot M^{(k;\alpha,\lambda)} \cdot (G^{(k,\lambda)})^{\alpha} \cdot a.$$

Hence, by B 3.2c, a necessary and sufficient condition for $(q_{(1)} \ldots q_{(\lambda)} \cdot G^{(k,\lambda)})^{\alpha} \cdot a = 0$ for all such tensors $q_{(1)}, \ldots, q_{(\lambda)}$ is

$$\{(G^{(k,\lambda)})^{\lambda} \cdot M^{(k;\alpha,\lambda)} \cdot (G^{(k,\lambda)})^{\alpha}\} \cdot a = 0.$$

A similar result also holds for Grassmann tensors, so that the result follows from A 4.1c and B 4.3f.

B 4.4 The intersection problem.

I shall now investigate the circumstances under which it is possible to express the coefficients of the Cayley form of the intersection of two varieties as polynomials in the coefficients of their Cayley forms.

This question can have several interpretations and one, at least, of these will admit a solution in all cases, However, I shall be concerned here with two more restrictive (and probably more natural) interpretations of the problem.

B 4.4a Notation. Let r, s, k, l, n-k-l $\geqslant$ 1. Then I shall use the symbol

$$\left\{ \begin{matrix} r & s \\ k & l \end{matrix} \right\}$$

to denote the following proposition :

"There exists a set of polynomials over $\mathfrak{A}$, in the coefficients of the Cayley forms of two unmixed multiplicative varieties f and g in $\mathfrak{B}$, of respective orders r and s and respective dimensions n-k-1 and n-l-1, which gives the set of coefficients of the Cayley form of their multiplicative intersection h, provided that this has dimension n-k-l-1, this same set of polynomials giving this required result for all such varieties f, g, h satisfying these conditions."

Similarly, I shall use a symbol $\left\{ \begin{matrix} r & .. & t \\ k & .. & m \end{matrix} \right\}$ to denote the corresponding proposition concerning the intersection of three or more multiplicative varieties.

B 4.4b The coefficients (in a frame $\mathcal{F}$) of the Cayley forms of the varieties f, g and h of B 4.4a are the components in $\mathcal{F}$ of the corresponding Cayley tensors f^H, g^H and h^H (except, perhaps, for certain numerical multipliers). Thus, the proposition $\left\{\begin{matrix} r & s \\ k & l \end{matrix}\right\}$ could equally well have been stated in terms of the components of f^H, g^H and h^H. It could also equally well have been stated in terms of f_G, g_G and h_G because B 3.5c, B 3.5e, B 3.5f, B 3.5g give the components of f_G as (linear) polynomials of the components of f^H, similarly for g_G and h_G, and conversely. Hence, an equivalent form of the proposition $\left\{\begin{matrix} r & s \\ k & l \end{matrix}\right\}$ is :

"There exists a set of polynomials over $\mathfrak{R}$, in the components in of f_G and g_G which gives the set of components in $\mathcal{F}$ of h_G, this result holding for all $f_G \in \mathcal{W}_G(k,r)$, $g_G \in \mathcal{W}_G(l,s)$, $h_G \in \mathcal{W}_G(k+l,rs)$ for which h is the intersection of f and g."

(The order of h must be rs by Bezout's theorem (see (10) pp.158-160.)

It may be thought that the conditions to be satisfied, in order that $\left\{\begin{matrix} r & s \\ k & l \end{matrix}\right\}$ be true for a particular r, s, k, l, are unnecessarily restrictive owing to the fact that the varieties $\mathcal{W}_G(k,r)$ and $\mathcal{W}_G(l,s)$ are often reducible. A possible situation would be the existence of a different set of polynomials for each pair of irreducible geometric components of $\mathcal{W}_G(k,r)$ and $\mathcal{W}_G(l,s)$. In order to cover this possibility, I shall introduce the following important concept.

B 4.4c Definition. Let a,..,c be non-zero elements of $\mathcal{C}$. Then I shall call the set (a',..,c') of elements of $\mathcal{C}$, a (tensor) specialization of (a,..,c) if a'..c'≠0 and

$$(a',..,c') \in \mathcal{Q}_{\{a,..,c\}} .$$

This is analogous to the idea of a proper specialization of a set of points (as given in (10) p.91), since $(a',..,c') \in \mathcal{Q}_{\{a,..,c\}}$ means that any relation of the form $(a)^{\lambda}..(c)^{\tau} \cdot A = 0$ where $A \in \mathfrak{R}$ implies that $(a')^{\lambda}..(c')^{\tau} \cdot A = 0$ also (see B 2.2d). Thus any geometric variety is the set of specializations of any one of its generic points. By B 2.2e and by B 2.2u, alternative conditions for (a',..,c') to be a specialization of (a,..,c), provided that a'..c' ≠ 0, are

$\mathcal{Q}_{\{a',..,c'\}} \subseteq \mathcal{Q}_{\{a,..,c\}}$, $E^{(\lambda,..,\tau)}_{\{a',..,c'\}} \cdot E^{(\lambda,..,\tau)}_{\{a,..,c\}} = E^{(\lambda,..,\tau)}_{\{a',..,c'\}}$ for each r,..,t $\geqslant$ 0, and

$(a')^{\lambda}..(c')^{\tau} \cdot E^{(\lambda,..,\tau)}_{\{a,..,c\}} = (a')^{\lambda}..(c')^{\tau}$.

B 4.4d Notation. Let $f_G \in \mathcal{W}_G(k,r)$ and $g_G \in \mathcal{W}_G(l,s)$, where r, s, k, l, n-k-l $\geqslant$ 1. Then the symbol

$$\{f, g\}$$

will be used to denote the following proposition :

"There exists a set of polynomials over $\mathfrak{R}$, in the components in a frame $\mathcal{F}$ of f_G and g_G, which gives the set of components in $\mathcal{F}$ of the Cayley tensor h_G of their multiplicative intersection, provided that this is purely (n-r-s-1)-dimensional, this same set of polynomials giving the corresponding result whenever f_G is replaced by a specialization f'_G of

f_ζ and g_ζ is replaced by a specialization g'_ζ of g_ζ if the intersection remains (n-r-s-1)-dimensional."

Let $\mathcal{W}'$ and $\mathcal{W}''$ be any irreducible geometric components of $\mathcal{W}_\zeta(k,r)$ and $\mathcal{W}_\zeta(l,s)$ respectively, (so that $\mathcal{W}' \times \mathcal{W}''$ is irreducible). By B 2.2g, it will be possible to choose f,g so that $\mathcal{W}' = \mathcal{U}_{\{f_\zeta\}}$ and $\mathcal{W}'' = \mathcal{U}_{\{g_\zeta\}}$, f_ζ and g_ζ being generic tensors of $\mathcal{W}'$ and $\mathcal{W}''$, respectively (see B 2.21). Hence, the possibility envisaged at the end of B 4.4b is covered by considerations of propositions of the type $\{f,g\}$.

B 4.4e If $\left\{\begin{matrix} r & s \\ k & l \end{matrix}\right\}$ is true then $\{f,g\}$ is true for all $f_\zeta \in \mathcal{W}_\zeta(k,r)$, $g_\zeta \in \mathcal{W}_\zeta(l,s)$.

B 4.4f <u>Notation</u>. Let r,..t, k,..,m, $n-(k+..+m)+1 \geqslant 1$. Then $\mathcal{X} = \mathcal{X}\left(\begin{matrix} \cdot & \cdot & t \\ k & \cdot & m \end{matrix}\right)$ denotes the subset of $\mathcal{W}_\zeta(k,r) \times .. \times \mathcal{W}_\zeta(m,t)$ consisting of the sets $(f_\zeta,..,h_\zeta)$ for which the intersection of the varieties of the sets f,..,h is not purely $(n-(k+..+m)-1)$-dimensional. If, in particular, $n = k+..+m$, then $(f_\zeta,..,h_\zeta) \in \mathcal{X}$ is the condition for the varieties f,..,h to have a point in common.

Now $|\mathcal{X}|$ is certainly a geometric variety. For suppose $k+..+m < n$ and let q be a dual Grassmann tensor corresponding to a $(k+..+m-1)$-space in $\mathfrak{P}$. Then the condition that $(f_\zeta,..,h_\zeta) \in \mathcal{X}$ is that for all such q's there exists an $x^\alpha \neq 0$ such that

$$q \cdot x I^{(n-k-..-m-1)} = 0,\ f_\zeta \cdot (x I^{(k-1)})^\lambda = 0,..,h_\zeta \cdot (x I^{(m-1)})^\chi = 0.$$

(That is, all such q's meet the intersection of f,..,h. See B 4.1d.) Hence, by B 3.7b, the condition is, that for all $q \in \mathcal{U}_{G_r}(k+..+m-1)$, the

tensors q, $f_G,..,h_G$ belong to a certain geometric variety. Thus, it follows from B 3.2c that $\mathscr{V}$ is a geometric variety. If k+..+m = n, the result follows simply from B 3.7b.

The notation $\mathscr{V}\left(\begin{smallmatrix}\lambda..t\\ k..m\end{smallmatrix}\middle|\begin{smallmatrix}\mu..w\\ a..c\end{smallmatrix}\right)$ will also be used for the corresponding geometric variety to $\mathscr{V}\left(\begin{smallmatrix}\lambda..t\,u..w\\ k..m\,a..c\end{smallmatrix}\right)$ but which is a subset of $\mathscr{W}_G(k,r) \times ..$ $\times \mathscr{W}_G(m,t) \times \mathscr{W}^G(a,u) \times .. \times \mathscr{W}^G(c,w)$.

B 4.4g The variety $\mathscr{V}\left(\begin{smallmatrix}\lambda..t\\ k..w\end{smallmatrix}\right)$ is certainly not the whole of $\mathscr{W}_G(k,r) \times .. \times \mathscr{W}_G(m,t)$ since the intersection of f ,..,h is "usually" (n-(k+..+m)-1)-dimensional. In fact, if $f_G,..,h_G$ are any elements of $\mathscr{W}_G(k,r),..,\mathscr{W}_G(m,t)$, respectively (where possibly $(f_G,..,h_G) \in \mathscr{V}$), then there exist tensors $f'_G,..,h'''_G$, each of which is the image, under a suitable linear transformation, of the corresponding $f_G,..,h_G$ where the intersection of $f',..,h'''$ is purely (n-(k+..+m)-1)-dimensional (see (10) pp.156–157). Since geometric varieties are invariant under linear transformations, $f'_G,..,h'''_G$ are specializations of $f_G..,h_G$, respectively.

B 4.4h Lemma. Let $f_G \in \mathscr{W}_G(k,r)$, $g_G \in \mathscr{W}_G(l,s)$, r, s, k, l, n-k-l $\geqslant$ 1, and suppose $\{f,g\}$ is true. Then there exist integers r', s', t and t' and an element A of $\mathscr{F}$ such that for each pair of specializations f'_G of f_G and g'_G of g_G for which $(f'_G,g'_G) \notin \mathscr{V}\left(\begin{smallmatrix}\lambda..\\ k..\end{smallmatrix}\right)$, a Cayley tensor h_G of the intersection of f' and g' is given by $h_G = (f'_G)^{\lambda'}(g'_G)^{\alpha'}(\epsilon^*)^t \cdot A \cdot G^{(\lambda..t,\alpha..)}(\epsilon)^{t'}$.

Proof. Without loss of generality $(f_G, g_G) \notin \mathscr{V}\left(\begin{smallmatrix}\lambda..\\ k..\end{smallmatrix}\right)$ (see B 4.4g). Then the components in the frame $\mathscr{F}$, of the tensor h_G corresponding to the intersection of f and g are given by polynomials in the components in $\mathscr{F}$

of f_σ and g_σ. Now if $\lambda, \mu \neq 0$ are scalars, then clearly λf_σ and μg_σ are tensor specializations of f_σ and g_σ. Thus, the polynomials are each altered by the same non-zero scalar factor when f_σ and g_σ are respectively replaced by λf_σ and μg_σ. It follows that the polynomials are separately homogeneous in the components of f_σ and of g_σ and are all of the same degrees.

Now, let p be a Grassmann tensor corresponding to an (1+k-1)-space in $\mathfrak{P}$. Then $h_\sigma \cdot (p)^{\wedge}$ is a form in the sets of components in $\mathcal{F}$ of h_σ and of p separately, whose vanishing is a necessary and sufficient condition for the space p to meet h. But we have seen that the components of h_σ are homogeneous polynomials in the components of f_σ and in the components of g_σ and are all of the same degrees. Hence $h_\sigma \cdot (p)^{\wedge} \equiv \phi(f_\sigma, g_\sigma, p)$ is also a form in the sets of components of each of f_σ, g_σ and p. The condition for p to meet the intersection of f and g is $\phi(f_\sigma, g_\sigma, p) = 0$ and this is to hold if f_σ and g_σ are replaced by any pair of specializations f'_σ and g'_σ for which $(f'_\sigma, g'_\sigma) \notin \mathfrak{D}^{(\hat{r}\hat{t})}$. In particular, it is to hold if $f'_\sigma = \Theta(f_\sigma)$ and $g'_\sigma = \Theta(g_\sigma)$ where Θ is some linear transformation of $\mathcal{C}$. It follows that the condition on the tensors f_σ, g_σ, p which is given by the vanishing of the polynomial $\phi(f_\sigma, g_\sigma, p)$, is independent of the particular choice of the frame in which the components of f_σ, g_σ, p are taken. Thus, by B 1.4 with $\mathfrak{Q} = \mathfrak{V}_{\{f_\sigma\}} \times \mathfrak{V}_{\{g_\sigma\}} \times \mathfrak{V}^\sigma(n-k-1)$, for all $f'_\sigma \in \mathfrak{V}_{\{f_\sigma\}}$, $g'_\sigma \in \mathfrak{V}_{\{g_\sigma\}}$ and $p \in \mathfrak{V}^\sigma(n-k-1)$, $\phi(f'_\sigma, g'_\sigma, p)$ is equal to a simple invariant of the

tensors f_Φ, g_Φ, p. That is $h_\Phi \bullet (p)^{\wedge\mu} \equiv \phi(f'_\Phi, g'_\Phi, p) \equiv (f'_\Phi)^{\wedge\nu'} (g'_\Phi)^{\wedge\sigma'} (\epsilon^*)^{\tau} \bullet A \bullet (p)^{\wedge\mu} (\epsilon)^{\tau'}$ for some $A \in \mathcal{P}$. This is an identity in p subject to $p \in \mathcal{U}^{\Phi}(n-k-1)$. Hence the result follows from B 3.2c.

B 4.4i Lemma. Let r, k, $n-k \geqslant 1$. Then for each $f_\Phi \in \mathcal{W}_\Phi(r,k)$ and each $q \in \mathcal{U}_\Phi(k)$ the tensor $(q)^{\wedge}$ is a specialization of f_Φ.

Proof. This result is essentially the well-known fact that any algebraic variety can be specialized (in the conventional sense) to a repeated linear space (see (HD) pp 141-5). However, it is a simple matter to give a direct proof, which is as follows.

Choose the Grassmann tensor $p^{\mu\wedge}$ so that the (k-1)-space it represents does not meet the variety f or the space q. Then we have $f_\Phi \bullet (p)^{\wedge} \neq 0$ and $q \bullet p \neq 0$. Hence, putting $e = I^{(n)}_{q} \bullet p I^{(n)}_{p}$, we have $e \neq 0$. By A4.41 and A 4.4n, there is a scalar $\lambda \neq 0$ such that $N^{(k)} \bullet (e)^{k} = \lambda\, pq$. Thus $f_\Phi \bullet (e)^{k\nu} = \{f_\Phi \bullet (N^{(k)})^{\wedge}\} \bullet (e)^{k\nu} = \{\lambda^{\wedge} f_\Phi \bullet (p)^{\wedge}\} (q)^{\wedge} = \mu (q)^{\wedge}$ where $\mu = \lambda^{\wedge} f_\Phi \bullet (p)^{\wedge} \neq 0$. Now suppose that the geometric variety $\mathcal{V}_{\{f_\Phi\}}$ is given by $(x)^m \bullet F = 0$ where $F \in \mathcal{T}_{0(n(n-1)\ldots k+1)m}$. Then we have

$$\{(q)^{\wedge}\}^m \bullet F = \mu^{-m} \{f_\Phi \bullet (e)^{k\nu}\}^m \bullet F = \mu^{-m} (f_\Phi)^m \bullet F \bullet (e)^{k\nu m} = 0$$

(by A 1.8u and by B 2.2c) so that $(q)^{\wedge} \in \mathcal{V}_{\{f_\Phi\}}$ as required.

This result also holds with f_Φ replaced by any tensor b satisfying $b = b \bullet G^{(k,\nu)} \neq 0$ as is readily seen from the proof.

B 4.4j Lemma. If f_Φ, g_Φ, f'_Φ, g'_Φ and h_Φ are as in B 4.4h, then A may be chosen so that $r' = s$, $s' = r$ and $t' = t = 0$.

Proof. The formula of B 4.4h is to hold for any specializations f'_ζ of f_ζ and g'_ζ of g_ζ for which $(f'_\zeta, g'_\zeta) \notin \mathfrak{F}$. Now let q and q′ be any elements of $\mathfrak{V}_\zeta(k)$ and $\mathfrak{V}_\zeta(l)$ respectively, for which $(q,q') \notin \mathfrak{F}\binom{11}{kl}$. By B4.4i , we can put $f'_\zeta = (q)^\wedge$ and $g'_\zeta = (q')^\wedge$, the intersection of f′ and g′ being the (n-k-m-1)-space of intersection of q and q′ counted rs times, so that $(f', g') \notin \mathfrak{F}\binom{\wedge\wedge}{kl}$.

Hence, $(q)^{\wedge\wedge'}(q')^{\wedge\wedge'}(\epsilon^*)^t \cdot A \cdot (p)^{\wedge\wedge} (\epsilon)^{t'} \neq 0$, for suitable p. Let $p = N^{(k+l)} \cdot x_{(1)} \ldots x_{(k+l)}$, $q = y_{(1)} \ldots y_{(k)} \cdot N^{(k)}$, $q' = y_{(k+1)} \ldots y_{(k+l)} \cdot N^{(l)}$. Now, we may suppose that either $t = 0$ or $t' = 0$ (by A 4.5b). Suppose first that $t = 0$, then

$$0 \neq (y_{(1)} \ldots y_{(k)} \cdot N^{(k)})^{\wedge\wedge'} (y_{(k+1)} \ldots y_{(k+l)} \cdot N^{(l)})^{\wedge\wedge'} \cdot A \cdot (N^{(k+l)} \cdot x_{(1)} \ldots x_{(k+l)})^{\wedge\wedge} (\epsilon)^{t'}.$$

Expanding $N^{(k)}$, $N^{(l)}$ and A as sums of permutations, we get

$$0 \neq \{ y_{(i_1)} \ldots y_{(i_{k+l})} \cdot N^{(k+l)} \cdot x_{(1)} \ldots x_{(k+l)} \} \ldots \{ y_{(i''_1)} \ldots y_{(i''_{k+l})} \cdot N^{(k+l)} \cdot x_{(1)} \ldots x_{(k+l)} \}$$
$$\{ y_{(j_1)} \ldots y_{(j_m)} \cdot \epsilon \} \ldots \{ y_{(j''_1)} \ldots y_{(j''_m)} \cdot \epsilon \}$$

for a suitable permutation $(i_1, \ldots, i_{k+l}, \ldots, i''_1, \ldots, i''_{k+l}, j_1, \ldots, j_m, \ldots, j''_1, \ldots j''_m)$ of $(1, \ldots, k, \ldots, 1, \ldots, k, k+1, \ldots, k+1, \ldots, k+1, \ldots, k+1)$.

But $n > k + 1$, so that any expression $y_{(j_1)} \ldots y_{(j_m)} \cdot \epsilon$ must vanish because of repetitions among the y's. Hence $t' = 0$.

Also, at least one of the expressions like $y_{(i_1)} \ldots y_{(i_{k+l})} \cdot N^{(k+l)} \cdot x_{(1)} \ldots x_{(k+l)}$ will vanish due to such repetitions unless each of $y_{(1)}, \ldots, y_{(k+l)}$ occurs exactly once in each factor. Hence $rr' = ss' = rs$, so that $r' = s$, $s' = r$.

By a similar argument it can be shown that if it is supposed that $t' = 0$

then it follows that $t = 0$, so that $r' = s$ and $s' = r$ as before.

(It would not be correct to deduce that $r' = s$ and $s' = r$ simply from the fact that the intersection of f' and g' possesses a Cayley tensor of the form $(qq')^{\wedge\lambda} \bullet G^{(\lambda,4)}$, since a scalar of proportionality which may involve q and q' has to be considered.)

B 4.4k **Lemma.** A necessary and sufficient condition for $\{f, g\}$ to be true is the existence of an element V of $\mathcal{V}$ such that, for all $f'_{\zeta} \in \mathcal{U}_{\{f_\zeta\}}$, $g'_{\zeta} \in \mathcal{U}_{\{g_\zeta\}}$, $p \in \mathcal{U}^{\zeta}_{(n-k-l)}$, the relation $(f'_{\zeta})^{\lambda} (g'_{\zeta})^{\mu} \bullet V \bullet (p)^{\nu\lambda} = 0$ holds if and only if p meets the intersection of f' and g'.

Proof. The condition is clearly sufficient since if $(f'_{\zeta}, g'_{\zeta}) \notin \mathcal{K}$, then $(f'_{\zeta})^{\lambda} (g'_{\zeta})^{\mu} \bullet V \bullet (p)^{\nu\lambda}$ is the Cayley form of the intersection of f' and g' in terms of Grassmann coordinates.

Now, if $\{f,g\}$ is true, it follows from B 4.4j that there exists $V \in \mathcal{V}$ such that, for all $f'_{\zeta} \in \mathcal{U}_{\{f_\zeta\}}$, $g'_{\zeta} \in \mathcal{U}_{\{g_\zeta\}}$ and $p \in \mathcal{U}^{\zeta}_{(n-k-l)}$ for which $(f'_{\zeta}, g'_{\zeta}) \notin \mathcal{K}$, the relation $(f'_{\zeta})^{\lambda} (g'_{\zeta})^{\mu} \bullet V \bullet (p)^{\nu\lambda} = 0$ holds if and only if f', g' and p have a point in common. If $(f'_{\zeta}, g'_{\zeta}) \in \mathcal{K}$, then in any case f', g' and p must have a point P in common. But the generic varieties f'' and g'' through P for which $f''_{\zeta} \in \mathcal{U}_{\{f_\zeta\}}$ and $g''_{\zeta} \in \mathcal{U}_{\{g_\zeta\}}$ have an $(n-k-l-1)$-dimensional intersection. Hence $(f'_{\zeta})^{\lambda} (g'_{\zeta})^{\mu} \bullet V \bullet (p)^{\nu\lambda} = 0$ in this case also.

B 4.4l It follows from B 4.4k that a necessary and sufficient condition for $\{f,g\}$ to be true is the existence of $V \in \mathcal{V}$ such that, for any

specializations f'_ζ of f_ζ and g'_ζ of g_ζ, the tensor $(f'_\zeta)^{\alpha}(g'_\zeta)^{\lambda}\bullet V$ is a Cayley tensor h_ζ of the intersection of f with g unless $(f_\zeta, g_\zeta) \in \mathfrak{D}$ in which case this tensor vanishes. Also, by B 2.3e, such a V must be unique up to proportionality if $\{f, g\}$ is true and $V = E^{(\alpha)}_{\{f_\zeta\}} E^{(\lambda)}_{\{g_\zeta\}} \bullet V \bullet G^{(k+l,\lambda)}$

B 4.4m Notation. If $f_\zeta \in \mathcal{W}_\zeta(k,r)$, $g_\zeta \in \mathcal{W}_\zeta(l,s)$; r, s, k, l $\geqslant$ 1 as before, but n-k-l = 0, then I define the proposition $\{f,g\}$ to mean

"there is an element V of $\mathcal{P}$ such that $(f'_\zeta)^{\alpha}(g'_\zeta)^{\lambda}\bullet V\bullet(\epsilon\ )^{\alpha\lambda} = 0$ is a necessary and sufficient condition for specializations f'_ζ of f_ζ and g'_ζ of g_ζ to represent varieties which have a point in common."

The condition $\{\hat{k}\,\hat{l}\}$ is then defined as the proposition :

"there exists an element V of $\mathcal{P}$ such that for all $f_\zeta \in \mathcal{W}_\zeta(k,r)$, $g_\zeta \in \mathcal{W}_\zeta(l,s)$, the condition $(f_\zeta)^{\alpha}(g_\zeta)^{\lambda}\bullet V\bullet(\epsilon\ )^{\alpha\lambda} = 0$ is necessary and sufficient for the varieties f and g to have a point in common."

With this definition, it is clear that B 4.4e still holds with n-k-l = 0.

The condition $\{\hat{k}\cdots\hat{l}\}$, when n = k+..+l, may also be defined in a similar way.

That the definition is a suitable interpretation of the condition $\{\hat{k}\,\hat{l}^{\alpha}\}$ as defined in B 4.4a, but with k + l = n, is made clear by the following result :

B 4.4n Lemma. If r, s, k, l, n-k-l+1 $\geqslant$ 1, then $\{\hat{k}\,\hat{l}^{\alpha}\}$ is equivalent to the proposition :

"there exists an element V of $\mathcal{T}$ such that for all $f_G \in \mathcal{W}_G(k,r)$, $g_G \in \mathcal{W}_G(l,s)$, $p \in \mathcal{U}^G(k+l)$, the condition $(f_G)^A (g_G)^{\wedge} \cdot V \cdot (p)^{\wedge\wedge} = 0$ is necessary and sufficient for the space p and the varieties f and g to have a point in common."

(If n = k + l, the space p is the whole of $\mathfrak{P}$ and p = ϵ .)

Proof. If n = k + l this is just B 4.4m. Suppose n > k + l. As in the case of $\{f, g\}$, $\{\hat{k}\,\hat{l}\}$ asserts the existence of a polynomial in the components of f_G, g_G and p, in a frame $\mathcal{F}$, which vanishes if and only if f, g and p have a point in common, for all $f_G \in \mathcal{W}_G(k,r)$, $g_G \in \mathcal{W}_G(l,s)$, $(f_G, g_G) \notin \mathcal{D}$ and $p \in \mathcal{U}^G(n-k-l)$. This polynomial may be used as the polynomial whose existence is asserted by $\{f, g\}$ and thus, as in B 4.4k, it may be written as $(f_G)^A (g_G)^{\wedge} \cdot V \cdot (p)^{\wedge\wedge}$. The result follows from B 4.4k.

B 4.4o If $\{\hat{k}\,\hat{l}\}$ is true, the tensor V is determined up to proportionality by the properties required of it in B 4.4n if $V = W_G^{(k,\wedge;A)} W_G^{(l,A;\wedge)} \cdot V \cdot G^{(k+l,\wedge\wedge)}$. For let V and V′ be two such tensors. Then by B 4.4l, $E^{(a)}_{\{f_G\}}\, E^{(\wedge)}_{\{g_G\}} \cdot (V' - V) \cdot G^{(k+l,\wedge\wedge)} = 0$. But, by B 2.2t, we have $W_G^{(k,\wedge;a)} = \left[E^{(a)}_{\{f_G\}} \,;\, f_G \in \mathcal{W}_G(k,r) \right]$ and similarly for $W_G^{(l,a;\wedge)}$. Hence the result follows from A 4.2c .

B 4.4p Remark. B 4.4m tells us that if $\{\hat{k}\,\hat{l}\}$ is true, then $\mathcal{D}(\hat{k}\,\hat{l} \,|\, {}^{1}_{m-k-l})$ is the complete intersection of $\mathcal{W}_G(k,r) \times \mathcal{W}_G(l,s) \times \mathcal{U}^G(k+l)$ with a primal in $\mathcal{C}_{\alpha^{k}} \times \mathcal{C}_{\alpha^{l}} \times \mathcal{C}_{\alpha^{m-(k+l)}}$. (In fact an invariant primal.)
The converse of this result is also true.

B 4.4q **Lemma.** Let Ψ be a projection on $\mathcal{C}$, of rank m, and let $q_{\alpha^{n-m}}$ be a dual Grassmann tensor of the base space of Ψ. Let r, k, m-k $\geqslant 1$ and suppose $f_{G} \in \mathcal{W}_{G}(k,r)$ is a Cayley tensor in $\mathcal{C}$ for which $(f_{G}, q) \in \mathfrak{X}$ Then $\Psi(f_{G})$ is a Cayley tensor of the system $\Psi(\mathcal{C})$, and it corresponds to the intersection of f with q. Conversely, any Cayley tensor f'_{G} of $\Psi(\mathcal{C})$ arises in this way from a Cayley tensor f_{G} of $\mathcal{C}$.

Proof. Let e be the idempotent in $\mathcal{C}$ which defines Ψ. Then $f'_{G} = \Psi(f_{G})$ $= f_{G} \bullet (e)^{k_r}$. Let p^{α^m} be a Grassmann tensor of the vertex space of Ψ Then, by B 1.5d, we have $N^{(k)} \bullet (e)^{k} = \mu q I^{(k)} \bullet N^{(n-m+k)} \bullet p I^{(k)}$ where $\mu \neq 0$. Now $f_{G} = f_{G} \bullet (N^{(k)})^{r}$, whence

$$f'_{G} = \mu^{\wedge} f_{G} \bullet \{ q I^{(k)} \bullet N^{(n-m+k)} \bullet p I^{(k)} \}^{\wedge} = f''_{G} \bullet (p I^{(k)})^{\wedge}$$

where $f''_{G} = \mu^{\wedge} f_{G} \bullet \{ q I^{(k)} \bullet N^{(n-m+k)} \}^{\wedge}$.

By B 4.1f, f''_{G} is a Cayley tensor in $\mathcal{C}$ of the intersection of f with q. Let x^{α} be any element of $\Psi(\mathcal{C}_{\alpha^*})$. Then x corresponds to a point of q. The tensor $N^{(n-m+1)} \bullet px$ is thus a Grassmann tensor of the join of the point x to the space p (see A 4.4f). Now since the spaces q and p are disjoint, the join of x to p meets the variety f'' when and only when x lies on f''. Thus, by B 4.1c, a necessary and sufficient condition for the point x of to lie on f'' is

$$f''_{G} \bullet \{ (N^{(n-m+1)} \bullet px) I^{(k-1)} \}^{\wedge} = 0 \ ;$$

that is, (by A 1.5u)

$$f''_{G} \bullet (px I^{(k-1)})^{\wedge} = 0$$

i.e. $$f'_{\zeta} \bullet (xI^{(k-1)})^{\lambda} = 0$$

Also, since $f_\zeta = f_\zeta \bullet G^{(k,\lambda)}$, it follows from B 1.5a that

$$f'_\zeta = f'_\zeta \bullet G^{(k,\lambda)} = f'_\zeta \bullet \{\Psi(G^{(k,\lambda)})\}$$

Hence, by B 4.1e, f'_ζ is a Cayley tensor of the system $\Psi(\mathcal{C})$ corresponding to the intersection of f with q, as required.

The converse is immediate, since any variety in $|\Psi(\mathcal{C}_{\alpha*})|$ arises as a suitable intersection with some variety in $|\mathcal{C}_{\alpha*}| = \mathfrak{F}_\beta$. For example, we may choose the cone projecting it from the vertex space p of Ψ.

B 4.4r Lemma. Let $f_\zeta \in \mathcal{W}_\zeta(k,r)$, $g_\zeta \in \mathcal{W}_\zeta(l,s)$, $q' \in \mathcal{U}^\zeta(n-k-l)$; r, s, k, l, $n-k-l \geqslant 1$. Let Ψ be a projection on $\mathcal{C}$ (of rank k+1) with base space q' and let $f'_\zeta = \Psi(f_\zeta)$, $g'_\zeta = \Psi(g_\zeta)$. Then if $V \in \mathfrak{J}_{((k+l)\lambda\mu)}$, the expression

$$(f_\zeta)^\lambda (g_\zeta)^\mu \bullet V \bullet (q')^{\nu\kappa}$$

vanishes if and only if the corresponding expression

$$(f'_\zeta)^\lambda (g'_\zeta)^\mu \bullet V \bullet (\epsilon^{\{\Psi\}})^{\nu\kappa}$$

vanishes, $\epsilon^{\{\Psi\}}$ being an α-alternating tensor in $\Psi(\mathcal{C})$.

Proof. See B 1.5e and B 1.5 a ($V \bullet (q')^{\nu\kappa} = \Psi(V \bullet (q')^{\nu\kappa}) = (e)^{(k+l)\nu\kappa} \bullet V \bullet \epsilon^{\{\Psi\}}$).

B 4.4s Lemma. If $\{\overset{\lambda\ \mu}{k\ l}\}$ is true for a particular set of values of r, s, k, l and of n (r, s, k, l, n-k-l+1 $\geqslant$ 1), then it is true independently of n. Also, the element V of $\mathfrak{J}$ arising from $\{\overset{\lambda\ \mu}{k\ l}\}$ in B 4.4o is independent of n.

Proof. If $f_\zeta \in \mathcal{W}_\zeta(k,r)$, $g_\zeta \in \mathcal{W}_\zeta(l,s)$, $q \in \mathcal{U}^\zeta_{(n-k-l)}$ and Ψ is a protection on with base space q, then the intersection of q with f meets the intersection of q with g if and only if the varieties in $|\Psi(\mathcal{C}_{\alpha*})|$ represented by $\Psi(f_\zeta)$ and

$\mathcal{V}(g_G)$ meet. Hence the result follows from B 4.4n, B 4.4g and B 4.4r. (The element V of B 4.4o determines a unique element of $\mathcal{Y}$ since $G^{(k+l,\lambda d)} \cdot Q_m^{((k+l)\lambda d)} = 0$ if $n \geqslant k+1$).

B 4.5 Theorem. Let r, s, k, l, n-k-l $\geqslant 1$ and suppose that $\{\hat{\lambda}^{q}_{k\,l}\}$ is true. Then for each $f_G \in \mathcal{W}_G(k,r)$, $g_G \in \mathcal{W}_G(l,s)$ for which $(f_G, g_G) \notin \mathcal{D}^{(\lambda)}_{k\,l}$, a Cayley tensor h_G of the multiplicative intersection of f and g is given by

$$h_G = (f_G)^{d} (g_G)^{\lambda} \cdot S^{(k,\lambda,d)} I^{(l\lambda\lambda)} \cdot D^{(k:l;\lambda d)} \cdot G^{(k+l,\lambda d)} .$$

Proof. It follows from B 4.4s (and B 4.4n) that it will be sufficient to verify that if n = k + 1, the validity of $\{\hat{\lambda}^{q}_{k\,l}\}$ implies that $(f_G)^{d} (g_G)^{\lambda} \cdot S^{(k,\lambda,d)} I^{(l\lambda\lambda)} \cdot D^{(k:l;\lambda d)} \cdot (\epsilon)^{\lambda d}$ vanishes for $f_G \in \mathcal{W}_G(k,r)$, $g_G \in \mathcal{W}_G(1,s)$ if and only if the varieties f and g have a point in common.

Now, by the definition of $D^{(\cdots)}$ (see A 4.5e), $D^{(k:l;\lambda d)} = D^{(k\lambda:l\lambda;\lambda)} \cdot (D^{(k:l;d)})^{\lambda}$, Hence the expression is equal to

$$(f_G)^{d} \cdot S^{(k,\lambda,d)} \cdot \left\{ I^{(l\lambda\lambda)} (g_G)^{\lambda} \right\} \cdot D^{(k\lambda:l\lambda;\lambda)} \cdot (D^{(k:l;d)})^{\lambda} \cdot (\epsilon)^{\lambda d}$$

$$= (f_G)^{d} \cdot S^{(k,\lambda,d)} \cdot \left\{ I^{(l\lambda\lambda)} g_G \cdot D^{(k:l;d)} \cdot (\epsilon)^{d} \right\}^{\lambda} = (s!)^{\lambda} (f_G)^{d} \cdot S^{(k,\lambda,d)} \cdot (g^G)^{\lambda}$$

by A 4.5f and B 3.5c. Thus theorem B 4.5 is a consequency of the following lemma:

B 4.5a Lemma. The proposition $\{\hat{\lambda}^{\lambda}_{k\;n-k}\}$ is true if and only if for each $f_G \in \mathcal{W}_G(k,r)$ and $g^G \in \mathcal{W}^G(n-k,s)$, the condition

$$(f_G)^{d} \cdot S^{(k,\lambda,d)} \cdot (g^G)^{\lambda} = 0$$

is necessary and sufficient for the varieties f and g to have a point in common.

Proof. If the condition $(f_\varphi)^\lambda \cdot S^{(\lambda,\mu,\nu)} \cdot (g_\varphi)^\mu = 0$ is necessary and sufficient for f to meet g, then reversing the above argument we see that $(f_\varphi)^\lambda (g_\varphi)^\mu \cdot V \cdot (\in)^{\nu\rho} = 0$ is also such a condition, where $V = S^{(\lambda,\mu,\nu)} I^{(\rho,\sigma,\tau)} D^{(\lambda,\mu,\nu;\rho,\sigma,\tau)}$. Thus, by B 4.4m, $\{{}^{\mu\ \nu}_{\lambda\ \rho}\}$ is true.

Conversely, if $\{{}^{\mu\ \nu}_{\lambda\ \rho}\}$ is true, then there is an element of $\mathcal{J}$ such that $(f_\varphi)^\lambda (g_\varphi)^\mu \cdot V \cdot (\in)^{\nu\rho} = 0$ is necessary and sufficient, for each $f_\varphi \in \mathcal{W}_\varphi(k,r)$ and $g_\varphi \in \mathcal{W}_\varphi(n-k,s)$, for f and g to have a point in common. Now, by B 3.5e, $g_\varphi = \frac{1}{\{(n-k)!\}^\rho} (\in^*)^\rho \cdot D^{(n-k;\rho,\sigma,\tau)} \cdot g^\varphi I^{(\rho,\sigma)}$ so that the condition can be put in the form $(f_\varphi)^\lambda (\in^*)^{\nu\rho} \cdot V' \cdot (\in)^{\rho\sigma} (g^\varphi)^\sigma = 0$ where $V' \in \mathcal{J}$, for $f_\varphi \in \mathcal{W}_\varphi(k,r)$, $g^\varphi \in \mathcal{W}^\varphi(n-k,s)$. Hence, using A 4.5b, we obtain the condition in the form $(f_\varphi)^\lambda \cdot X \cdot (g^\varphi)^\sigma = 0$ with $X \in \mathcal{J}$. By B 4.2c, we can suppose, without loss of generality that $X = W_\varphi^{(\lambda,\mu,\nu)} \cdot X \cdot W^{\varphi(\lambda,\mu,\nu)}$. Now, if the condition $(f_\varphi)^\lambda \cdot X \cdot (g^\varphi)^\sigma = 0$ is necessary and sufficient, for each $f_\varphi \in \mathcal{W}_\varphi(k,r)$, $g^\varphi \in \mathcal{W}^\varphi(n-k,s)$, for the varieties f and g to meet, then a fortiori it is necessary and sufficient if $f_\varphi \in \mathcal{U}_\varphi(k,r)$ and $g^\varphi \in \mathcal{U}^\varphi(n-k,s)$ (see B 4.3a).

Hence B 4.5a is a consequence of the following two lemmas.

B 4.5b **Lemma.** If r, s, k, n-k $\geqslant 1$, then for each $f_\varphi \in \mathcal{U}_\varphi(k,r)$ and $g^\varphi \in \mathcal{U}^\varphi(n-k,s)$ the relation

$$(f_\varphi)^\lambda \cdot S^{(\lambda,\mu,\nu)} \cdot (g^\varphi)^\sigma = 0$$

is necessary and sufficient for the varieties f and g to meet. Also,

$$S^{(\lambda,\sigma,\nu)} = U^{(\lambda,\mu,\rho)} \cdot S^{(\rho,\mu,\nu)} \cdot U^{(\rho,\sigma,\nu)}$$

and the tensor $S^{(k,\lambda,\Delta)}$ is uniquely determined, except for a scalar of proportionality, by these two properties.

B 4.5c <u>Lemma</u>. If $\left\{\begin{matrix}\lambda & \Delta \\ k & n-k\end{matrix}\right\}$ is true, then $W_G^{(k,\lambda;\Delta)} = U^{(k,\lambda,\Delta)}$ and $W^{G(k,\Delta,\lambda)} = U^{(k,\Delta,\lambda)}$.

<u>Proof of B 4.5b</u>. The assumption of the validity of $\left\{\begin{matrix}\lambda & \Delta \\ k & n-k\end{matrix}\right\}$ is not needed for this result. Now the elements f_ς of $\mathcal{U}_{\varsigma}(k,r)$ and g^G of $\mathcal{U}^G(n-k,s)$ are respectively of the form

$$f_\varsigma = q_{(1)} \cdot\cdot q_{(\lambda)} \cdot G^{(k,\lambda)} \text{ and } g^G = G^{(k,\Delta)} \cdot p_{(1)} \cdot\cdot p_{(\Delta)}$$

where $q_{(1)},\ldots,q_{(\lambda)} \in \mathcal{U}_{\varsigma}(k)$ and $p_{(1)},\ldots,p_{(\Delta)} \in \mathcal{U}^G(n-k)$. (See B 4.3) The condition for the varieties f and g to meet is that at least one of the spaces $q_{(1)},\ldots,q_{(\lambda)}$ meet at least one of $p_{(1)},\ldots,p_{(\Delta)}$. That is, by A 4.4g , the product of $q_{(1)} \cdot p_{(1)},\ldots,q_{(1)} \cdot p_{(\Delta)}$, ..., $q_{(\lambda)} \cdot p_{(1)},\ldots,q_{(\lambda)} \cdot p_{(\Delta)}$ must vanish. Put $q = (q_{(1)})^\Delta \cdot\cdot (q_{(\lambda)})^\Delta$, $p = (p_{(1)})^\lambda \cdot\cdot (p_{(\Delta)})^\lambda$, $M = M^{(k;\Delta,\lambda)}$, $G = (G^{(k,\Delta)})^\lambda$, $G' = (G^{(k,\lambda)})^\Delta$ and $S = S^{(k,\lambda,\Delta)}$. Then this product is equal to

$$\begin{aligned} q \cdot (p_{(1)} \cdot\cdot p_{(\Delta)})^\lambda &= q \cdot M \cdot p = q \cdot G \cdot M \cdot G' \cdot p = q \cdot (G \cdot M \cdot G') \cdot S \cdot (G \cdot M \cdot G') \cdot p \\ &= q \cdot M \cdot G' \cdot S \cdot G \cdot M \cdot p = (q_{(1)} \cdot\cdot q_{(\lambda)})^\Delta \cdot G' \cdot S \cdot G \cdot (p_{(1)} \cdot\cdot p_{(\Delta)})^\lambda \\ &= (f_\varsigma)^\Delta \cdot S \cdot (g^G)^\lambda \end{aligned}$$

by B 3.1d, B 3.2c and B 4.3g. Thus, the condition $(f_\varsigma)^\Delta \cdot S^{(k,\lambda,\Delta)} \cdot (g^G)^\lambda = 0$ is necessary and sufficient for the variety f to meet the variety g as required. Also, since by B 4.3h (and A 2.2j), $U^{(k,\lambda,\Delta)} = S^{(k,\lambda,\Delta)} \cdot S^{(k,\lambda,\Delta)\dagger}$ and $U^{(k,\Delta,\lambda)} = S^{(k,\lambda,\Delta)\dagger} \cdot S^{(k,\lambda,\Delta)}$, we have $U^{(k,\lambda,\Delta)} \cdot S^{(k,\lambda,\Delta)} \cdot U^{(k,\Delta,\lambda)} = S^{(k,\lambda,\Delta)}$.

The uniqueness property of $S^{(k,\lambda,\Delta)}$ follows from B 4.4o.

<u>Proof of B 4.5c</u>. We have seen that if $\left\{\begin{matrix}\lambda & \Delta \\ k & n-k\end{matrix}\right\}$ is true, then for some $X \in \mathcal{F}$, the condition $(f_\varsigma)^\Delta \cdot X \cdot (g^G)^\lambda = 0$ is necessary and sufficient, for all

$f_G \in \mathscr{W}_G(k,r)$, $g^G \in \mathscr{W}^G(n-k,s)$ for the varieties f and g to meet. In particular, this must be true if $g^G \in \mathscr{U}^G(n-k,s)$. But then g^G is of the form $g^G = G^{(\mathfrak{h},\lambda)} \bullet p_{(1)} \ldots p_{(\lambda)}$ where $p_{(1)}, \ldots, p_{(\lambda)} \in \mathscr{U}^G(n-k)$, and the condition for g to meet f is the condition that one, at least, of $p_{(1)}, \ldots, p_{(\lambda)}$ should meet f. Now, the condition that $p_{(i)}$ meet f is $f_G \bullet (p_{(i)})^\lambda = 0$ (see B 4.1). Thus, a necessary and sufficient condition for g to meet f is that the product $\{f_G \bullet (p_{(1)})^\lambda\} \ldots \{f_G \bullet (p_{(\lambda)})^\lambda\}$ should vanish, i.e. that

$$(f_G)^\lambda \bullet \{(p_{(1)})^\lambda \ldots (p_{(\lambda)})^\lambda\} = 0.$$

But $(f_G)^\lambda \bullet X \bullet (g^G)^\lambda = (f_G)^\lambda \bullet X \bullet \{G^{(\mathfrak{h},\lambda)} \bullet p_{(1)} \ldots p_{(\lambda)}\}^\lambda$

$$= (f_G)^\lambda \bullet X \bullet (G^{(\mathfrak{h},\lambda)})^\lambda \bullet M^{(\mathfrak{h};\lambda,\lambda)} \bullet (p_{(1)})^\lambda \ldots (p_{(\lambda)})^\lambda ,$$

and the vanishing of this is also necessary and sufficient for g to meet f. Hence the subvariety of $\mathscr{U}^G(n-k) \times \ldots \times \mathscr{U}^G(n-k)$ given by

$$(f_G)^\lambda \bullet \{(x_{(1)})^\lambda \ldots (x_{(\lambda)})^\lambda\} = 0$$

is the same as that given by

$$\{(f_G)^\lambda \bullet X \bullet (G^{(\mathfrak{h},\lambda)})^\lambda \bullet M^{(\mathfrak{h};\lambda,\lambda)}\} \bullet \{(x_{(1)})^\lambda \ldots (x_{(\lambda)})^\lambda\} = 0.$$

Also, these subvarieties must be the same multiplicative intersections of the respective primals with $\mathscr{U}^G(n-k) \times \ldots \times \mathscr{U}^G(n-k)$. For if f has no repeated components the multiplicities are all unity, whereas if f has repeated components, f is a specialization of a variety without repeated components.

Thus, by B 2.3e (and B 2.3d, B 3.2b) we get

$$(f_G)^\lambda \bullet (G^{(\mathfrak{h},\lambda)})^\lambda = \lambda (f_G)^\lambda \bullet X \bullet (G^{(\mathfrak{h},\lambda)})^\lambda \bullet M^{(\mathfrak{h};\lambda,\lambda)} \bullet (G^{(\mathfrak{h},\lambda)})^\lambda$$

for some $\lambda \in \mathfrak{R}$. But $f_G \bullet G^{(\mathfrak{h},\lambda)} = f_G$ and

$$\{(G^{(\lambda,d)})^{\lambda}\cdot M^{(\lambda;d,\nu)}\cdot(G^{(\lambda,\nu)})^{d}\}\cdot U^{(\lambda,\nu,d)} = (G^{(\lambda,d)})^{\lambda}\cdot M^{(\lambda;d,\nu)}\cdot(G^{(\lambda,\nu)})^{d}$$

by B 4.3h and B 4.3g. Hence $(f_G)^{\lambda}\cdot U^{(\lambda,\nu,d)} = (f_G)^{\lambda}$.

This must hold for all $f_G \in \mathcal{W}_G(k,r)$ so that, by B 4.2d, we get

$$W_G^{(\lambda,\nu;d)}\cdot U^{(\lambda,\nu,d)} = W_G^{(\lambda,\nu;d)}$$

But $\mathcal{U}_G(k,r) \subseteq \mathcal{W}_G(k,r)$ so that, by B 2.2u (and A 4.1e), we have

$$W_G^{(\lambda,\nu;d)}\cdot U^{(\lambda,\nu,d)} = U^{(\lambda,\nu,d)}$$

proving $W_G^{(\lambda,\nu;d)} = U^{(\lambda,\nu,d)}$ as required.

In an exactly similar way, by considering the case $f_G \in \mathcal{U}_G(k,r)$, $g^G \in \mathcal{W}^G(n-k,r)$ we also obtain $W^{G(\lambda,d;\nu)} = U^{(\lambda,d,\nu)}$.

This completes the proof of the theorem.

B 4.5d Remark. The key to theorem B 4.5 is Lemma B 4.5c. For it shows that when $\{\hat{\lambda}\,\hat{d}\}$ is true, the problem of finding the intersection of two general varieties can be reduced to the special case in which both varieties consist solely of linear spaces.

Lemma B 4.5c has another interesting consequence. It shows that there are values of r, s, k, l for which $\{\hat{\lambda}\,\hat{d}\}$ is not true. The varieties $\mathcal{W}_G(k,r)$ and $\mathcal{U}_G(k,r)$ are certainly not always identical. (They are distinct if $n-k \neq 1 \neq r$). Hence, for such r, k, the corresponding projectors $W_G^{(\lambda,\nu;d)}$ and $U^{(\lambda,\nu,d)}$ must differ for some value of s (see B 2.4g).

It seems at least possible that the converse of Lemma B 4.5c is true, however (that is, $W_G^{(\lambda,\nu;d)} = U^{(\lambda,\nu,d)}$ and $W^{G(\lambda,d;\nu)} = U^{(\lambda,d,\nu)}$ together imply $\{\hat{\lambda}\,\hat{d}\}$). That this is so if k = 1 (the case of a primal and a set of points) will

follow from the next lemma, since by B 4.3c,

$$W^{G(1,d;\lambda)} = U^{(1,d,\lambda)}$$

for each s, r $\geqslant$ 1, and also

$$W_G^{(1,\lambda;d)} = P^{(\lambda,d)}_{\bullet}(P^{(\lambda)})^d$$

since the elements of $\mathscr{W}_G(1,r)$ are just the symmetric elements of $\mathcal{C}_{\alpha^\lambda}$.

B 4.5e **Lemma.** Let r, s $\geqslant$ 1. Then a necessary and sufficient condition for $\{^{\lambda_1 \ldots d}_{1 \ldots m-1}\}$ to be true is

$$U^{(1,\lambda,d)} = P^{(\lambda,d)}_{\bullet}(P^{(\lambda)})^d .$$

Proof. The necessity of the condition follows from B 4.5c since $W_G^{(1,\lambda;d)} = P^{(\lambda,d)}_{\bullet}(P^{(\lambda)})^d$ (as above).

Conversely, suppose that $U^{(1,\lambda,d)} = P^{(\lambda,d)}_{\bullet}(P^{(\lambda)})^d$ and let $f_G \in \mathscr{W}_G(1,r)$ and $g^G \in \mathscr{W}^G(n-1,r)$. Then, by B 4.5a, the result will follow if it can be shown that the relation

$$(f_G)^d \bullet S^{(\lambda,d)}_{\bullet}(g^G)^\lambda = 0$$

holds if and only if f meets g. ($S^{(\lambda,d)} = S^{(1,\lambda,d)}$, see B 4.3g.) Now, g is a set of points ($\mathscr{W}^G_{(n-1,\lambda)} = \mathscr{U}^G_{(n-1,\lambda)}$), i.e. g^G has the form

$$g^G = P^{(d)} \bullet x_{(1)} \ldots x_{(d)}$$

where $x^\alpha_{(1)}, \ldots, x^\alpha_{(d)} \in \mathcal{C}_{\alpha^*}$ ($G^{(1,d)} = P^{(d)}$). Hence

$$\begin{aligned}
(f_G)^d \bullet S^{(\lambda,d)}_{\bullet}(g^G)^\lambda &= (f_G)^d \bullet S^{(\lambda,d)}_{\bullet}(P^{(d)})^\lambda \bullet (x_{(1)} \ldots x_{(d)})^\lambda \\
&= (f_G)^d \bullet S^{(\lambda,d)}_{\bullet}(P^{(d)})^\lambda \bullet M^{(d,\lambda)}_{\bullet}\{(x_{(1)})^\lambda \ldots (x_{(d)})^\lambda\} \\
&= (f_G)^d \bullet S^{(\lambda,d)}_{\bullet}(P^{(d)})^\lambda \bullet M^{(d,\lambda)}_{\bullet}(P^{(\lambda)})^d \bullet \{(x_{(1)})^\lambda \ldots (x_{(d)})^\lambda\} \\
&= (f_G)^d \bullet U^{(1,\lambda,d)}_{\bullet}\{(x_{(1)})^\lambda \ldots (x_{(d)})^\lambda\}
\end{aligned}$$

$$= (f_G)^{\lambda} \bullet P^{(\lambda,q)} \bullet (P^{(\lambda)})^{q} \bullet \{(x_{(1)})^{\lambda} .. (x_{(q)})^{\lambda}\}$$

$$= (f_G)^{\lambda} \bullet \{(x_{(1)})^{\lambda} .. (x_{(q)})^{\lambda}\} = \{f_G \bullet (x_{(1)})^{\lambda}\} .. \{f_G \bullet (x_{(q)})^{\lambda}\}$$

But the condition that f should meet g is simply that f should contain at least one of the points $x_{(1)}, .., x_{(q)}$, i.e. that at least one of the expressions $f_G \bullet (x_{(i)})^{\lambda}$, for i = 1,..,s, should vanish, that is,

$$\{f_G \bullet (x_{(1)})^{\lambda}\} .. \{f_G \bullet (x_{(q)})^{\lambda}\} = 0.$$

This proves the result.

B 4.5f **Lemma.** Let r, s, l n-l $\geqslant$ 1. Then a necessary and sufficient condition for $\{\overset{\lambda}{l}\overset{q}{r}\}$ to be true is that any polynomial of degree at most s in the coefficients of an (l+1)-ary form of order r, which vanishes whenever the form factorizes into r linear parts, should be identically zero.

Proof. We may suppose n = l+1 (by B 4.4s). Let the form be $a \bullet (x)^{\lambda}$ (in some frame $\mathcal{F}$,) where $a_{\alpha^{\lambda}}$, $x^{\alpha} \in \mathcal{C}$, $a = a \bullet P^{(\lambda)} \neq 0$. Without loss of generality, the polynomial is homogeneous of degree s and can thus be written as $(a)^{q} \bullet b$ (in $\mathcal{F}$) where $b^{\alpha^{\lambda q}} \in \mathcal{C}$.

Now the necessary and sufficient condition for $(a)^{q} \bullet b$ to vanish identically in symmetric tensors a (i.e. for $a \in \mathcal{W}_G(1,r)$) is that

$$P^{(\lambda,q)} \bullet (P^{(\lambda)})^{q} \bullet b = 0,$$

and this property determines the projector $P^{(\lambda,q)} \bullet (P^{(\lambda)})^{q}$ uniquely (see B2.3, etc.). Similarly, the necessary and sufficient condition for $(a)^{q} \bullet b$ to vanish for all a, for which the form $a \bullet (x)^{\lambda}$ is a product of r linear factors

(i.e. for which $a \in \mathscr{U}_G(1,r)$), is that

$$U^{(1,\lambda,\mu)} \cdot b = 0,$$

and this property determines the projector $U^{(1,\lambda,\mu)}$ uniquely (see B 4.3f). Thus, the result follows from B 4.5e.

B 4.5g Lemma. Let r, s, k, l, m, n-k-l-m $\geqslant 1$. Let $f_G \in \mathscr{W}_G(k,r)$, $g'_G \in \mathscr{W}_G(l,s)$, $q \in \mathscr{U}_G(m)$, $g_G \in \mathscr{W}_G(l+m,s)$ where g is the intersection of the space q with g , and where $(f_G, g_G) \notin \mathscr{D}(\hat{k}\, \hat{l+m})$ (so that $(f_G, g'_G) \notin \mathscr{D}(\hat{k}\, \hat{l})$ also). Then a necessary and sufficient condition for the multiplicative intersection h′ of f and g′ to be given by

$$h'_G = (f_G)^{\mu} (g'_G)^{\lambda} \cdot S^{(k,\lambda,\mu)} I^{(l+\mu)} \cdot D^{(k:l;\lambda\mu)} \cdot G^{(k+l,\lambda\mu)}$$

is that for all such q, the multiplicative intersection h of f and g is given by

$$h_G = (f_G)^{\mu} (g_G)^{\lambda} \cdot S^{(k,\lambda,\mu)} I^{((l+m)\lambda\mu)} \cdot D^{(k:l+m;\lambda\mu)} \cdot G^{(k+l+m,\lambda\mu)}.$$

Proof. Let $p' \in \mathscr{U}^G(n-k-l)$ where the space p′ is contained in the space q. Then, by A4.4h , p′ is of the form

$$p' = I^{(k+l)} q \cdot p$$

where $p \in \mathscr{U}^G(n-k-l-m)$, the space p′ being the intersection of p and q. Now, by B 4.1f, g_G can be taken to be given by

$$g_G = g'_G \cdot (I^{(l)} q \cdot N^{(l+m)})^{\mu}.$$

Hence,

$$(f_G)^{\mu} (g_G)^{\lambda} \cdot S^{(k,\lambda,\mu)} I^{((l+m)\lambda\mu)} \cdot D^{(k:l+m;\lambda\mu)} \cdot (p)^{\lambda\mu}$$
$$= (f_G)^{\mu} (g'_G)^{\lambda} \cdot S^{(k,\lambda,\mu)} (I^{(l)} q \cdot N^{(l+m)})^{\mu\lambda} \cdot D^{(k:l+m;\lambda\mu)} \cdot (p)^{\lambda\mu}$$

$$= (f_G)^{\lambda} (g'_G)^{\lambda} \bullet S^{(k,\lambda,\lambda)} I^{(l,\lambda,\lambda)} \bullet D^{(k:l;\lambda\lambda)} \bullet (I^{(k,l)} q)^{\lambda\lambda} \bullet (p)^{\lambda\lambda}$$

$$= (f_G)^{\lambda} (g'_G)^{\lambda} \bullet S^{(k,\lambda,\lambda)} I^{(l,\lambda,\lambda)} \bullet D^{(k:l;\lambda\lambda)} \bullet (p')^{\lambda\lambda} .$$

But the space p meets the intersection of f and g if and only if the space p' meets the intersection of f and g'. The result follows (since p' can be chosen arbitrarily from $\mathcal{U}^{G}$ $(n-k-1)$ with suitable q).

B 4.5h Lemma. If r, s, k, l, m, $n-k-l-m+1 \geqslant 1$ and $\left\{ {\lambda\ \lambda \atop k\ l,m} \right\}$ is true, then $\left\{ {\lambda\ \lambda \atop k\ l} \right\}$ is true.

Proof. If $\left\{ {\lambda\ \lambda \atop k\ l,m} \right\}$ is true and f_G, g_G and g'_G are as in B 4.5g, then, by B 4.5, the intersection h of f and g is given as in B 4.5g. Hence the intersection h' of f and g' is also as given. But this holds for all $f_G \in \mathcal{W}_G(k,r)$, $g'_G \in \mathcal{W}_G(l,s)$, so that $\left\{ {\lambda\ \lambda \atop k\ l} \right\}$ is true, as required.

B 4.5i Lemma. If $\left\{ {\lambda\ \lambda \atop 1\ \lambda-1} \right\}$ is true, then $\left\{ {\lambda\ \lambda \atop 1\ l} \right\}$ is true for all l.

Proof. By B 4.5e, a necessary and sufficient condition for $\left\{ {\lambda\ \lambda \atop 1\ n-1} \right\}$ to be true is

$$S^{(\lambda,\lambda)}_{(\alpha)} \bullet \left\{ (P^{(\lambda)}_{(\mu)})^{\lambda} \bullet M^{(\lambda,\lambda)}_{(\alpha)} \bullet (P^{(\lambda)}_{(\alpha)})^{\lambda} \right\} = P^{(\lambda,\lambda)}_{(\alpha)} \bullet (P^{(\lambda)}_{(\alpha)})^{\lambda} .$$

But this is of the form $A_{(\alpha)} \bullet (P^{(\lambda)}_{(\alpha)})^{\lambda} = 0$ and is thus a substitutional identity if it is true when $n = s$ (by A 1.9d). Thus, if it is true when $n = s$, it is also true when n takes the value $l+1$. Hence, the result follows from B 4.4s.

B 4.5j Lemma. If $n = 2$ then $S^{(\lambda,\lambda)}_{(\alpha)} \bullet \left\{ (P^{(\lambda)}_{(\alpha)})^{\lambda} \bullet M^{(\lambda,\lambda)}_{(\alpha)} \bullet (P^{(\lambda)}_{(\alpha)})^{\lambda} \right\} = P^{(\lambda,\lambda)} \bullet (P^{(\lambda)})^{\lambda}$ for each $\lambda, \lambda \geqslant 1$.

Proof. This follows from B 4.5e (and B 4.3h) if $\left\{ {\lambda\ \lambda \atop 1\ 1} \right\}$ is true. But, as is

clear for a variety of reasons, $\{^{\lambda}_{1}{}^{\lambda}_{1}\}$ is always true. In particular, it is a consequence of the following result. The tensor $R^{(\lambda,..,t)}$ is as defined in B 3.7d.

B 4.6 Theorem. Let $r,..,t, n-(r+..+t)+1 \geqslant 1$. Then $\{^{\lambda..t}_{1..1}\}$ is true and, if $n > r+..+t$, then, for each $a \in \mathcal{W}_G(1,r), .., c \in \mathcal{W}_G(1,t)$, a Cayley tensor f_G of the intersection of the primals a,..,c is given by

$$f_G = \{(a)^{q/\lambda} .. (c)^{q/t}\} \cdot R^{(\lambda,..,t)}$$

where $q = r..t$ and $R^{(\lambda,..,t)}$ is as in B 3.7d, provided that $(a,..,c) \notin \mathcal{D}(^{\lambda..t}_{1..1})$

Proof. We saw in B 3.7d that if $n = r+..+t$, then the condition

$$\{(a)^{q/\lambda} .. (c)^{q/t}\} \cdot R^{(\lambda,..,t)} \cdot (\ \in\ )^q = 0$$

is necessary and sufficient for the primals a,..,c to have a point in common. Now the arguments leading up to and including B 4.4s will clearly also apply to conditions of the type $\{^{\lambda..t}_{k..m}\}$. Hence by B 4.4m and the appropriate modification of B 4.4s to several varieties, we see that $\{^{\lambda..t}_{1..1}\}$ is true (whether $n = r+..+t$ or not) and that if $n > r+..+t$ the condition

$$\{(a)^{q/\lambda} .. (c)^{q/t}\} \cdot R^{(\lambda,..,t)} \cdot (p)^q = 0$$

where $p \in \mathcal{G}^G(n-(r+..+t))$ is necessary and sufficient for the space p to meet the intersection of the primals a,..,c. The result follows.

B 4.6a Lemma. Let $r, s \geqslant 1$. Then

$$P^{(\lambda)} P^{(\lambda_4-\lambda)} \cdot S^{(\lambda,4)} \cdot \{(P^{(4)})^{\lambda} \cdot M^{(4,\lambda)} \cdot (P^{(\lambda)})^4\} = P^{(\lambda)} P^{(\lambda_4-\lambda)} \cdot P^{(\lambda,4)}$$

is a substitutional identity.

Proof. Since this is an identity of the form $P^{(k)}P^{(l)}\cdot A = 0$, it will be sufficient to prove that it is true for elements of $\mathcal{J}$ when n = 2, by A 1.9d. But $P^{(\lambda)}P^{(\lambda d-\lambda)}\cdot P^{(\lambda,d)} = P^{(\lambda)}P^{(\lambda d-\lambda)}\cdot(P^{(\lambda)})^d\cdot P^{(\lambda,d)} = P^{(\lambda)}P^{(\lambda d-\lambda)}\cdot P^{(\lambda,d)}\cdot(P^{(\lambda)})^d$ by A 1.5r and A 1.8u, so that the result is immediate from B 4.5j.

B 4.6b Lemma. Let $r, s \geqslant 1$ and $x_{(1)}^{\alpha},\ldots,x_{(d)}^{\alpha} \in \mathcal{C}$. Then

$$P^{(\lambda)}P^{(\lambda d-\lambda)}\cdot S^{(\lambda,d)}\cdot\left\{P^{(d)}\cdot x_{(1)}\ldots x_{(d)}\right\}^{\lambda} = \frac{1}{d}\sum_{i=1}^{d}(x_{(i)})^{\lambda}\left\{P^{(\lambda d-\lambda)}\cdot(x_{(1)})^{\lambda}\ldots(x_{(i-1)})^{\lambda}(x_{(i+1)})^{\lambda}\ldots(x_{(d)})^{\lambda}\right\}.$$

Proof. $\left\{P^{(d)}\cdot x_{(1)}\ldots x_{(d)}\right\}^{\lambda} = (P^{(d)})^{\lambda}\cdot M^{(d,\lambda)}\cdot(P^{(\lambda)})^d\cdot\left\{(x_{(1)})^{\lambda}\ldots(x_{(d)})^{\lambda}\right\}$ as in B 4.5e and $P^{(\lambda,d)}\cdot\left\{(x_{(1)})^{\lambda}\ldots(x_{(d)})^{\lambda}\right\} = \frac{1}{d}\sum_{i=1}^{d}(x_{(i)})^{\lambda}\left\{P^{(\lambda,d-1)}\cdot(x_{(1)})^{\lambda}\ldots(x_{(i-1)})^{\lambda}(x_{(i+1)})^{\lambda}\ldots(x_{(d)})^{\lambda}\right\}$ by A 1.5s. Hence the result follows from B 4.6a (and A 1.8u).

B 4.7 Theorem. The proposition $\left\{\begin{smallmatrix}\lambda & d\\ k & l\end{smallmatrix}\right\}$ is true in each of the following cases:

$k = l = 1$, B 4.7a

$s = 1$, B 4.7b

$r = s = 2$, B 4.7c

$s = 2$ and $k = 1$, B 4.7d

$s = 3$ and $k = 1$ but $r \neq 2$, B 4.7e

and it is false for each of the cases :

$k = 1$ and $r < s$ but $r \neq 1 \neq l$ B 4.7f

$r, s, k, l \neq 1$ and $r \neq s$. B 4.7g

Proof. The case B 4.7a has already been considered in B 4.6, and the case B 4.7b follows from B 4.1f. To prove B 4.7c, it will suffice to show that, for all $f_G \in \mathcal{W}_G(k,2)$, $g^G \in \mathcal{W}^G(1,2)$, where n = k+1, the condition

$$(f_\zeta)^2 \cdot S^{(k,2,2)} \cdot (g^\zeta)^2 = 0$$

is necessary and sufficient for the varieties f and g to meet (by B 4.5a).

Now the varieties of order 2 are of two kinds, namely the varieties which consist of two linear spaces of equal dimension (possibly skew or perhaps coincident) and the irreducible quadrics. Suppose, first, that both of f and g are reducible. Then the result follows from B 4.5b. Now suppose that at least one of f and g is irreducible. Without loss of generality g is irreducible, so that g is the intersection of a quadric primal g' with a linear space. As in the proof of B 4.5, it will be sufficient to show that for such a g, and where $n > k+1$, the expression

$$(f_\zeta)^2 (g_\zeta)^2 \cdot S^{(k,2,2)} I^{(\ell,\lambda)} \cdot D^{(k;\ell;\lambda)} \cdot G^{(k+\ell,\lambda)}$$

is a Cayley tensor of the intersection of f_ζ and g_ζ if $(f_\zeta, g_\zeta) \in \mathfrak{B}$. But, by B 4.5g, this will be so if a Cayley tensor of the intersection of g' with f is given as in B 4.5g. Thus the result will follow from $\{^{2\,2}_{k\,1}\}$ which is part of B 4.7d.

To prove B 4.7d, we observe that, by B 4.5i, $\{^{\lambda\,2}_{1\,\ell}\}$ follows from $\{^{\lambda\,2}_{1\,1}\}$ which is true by B 4.7a.

The case B 4.7e may be reduced to $\{^{\lambda\,r}_{1\,2}\}$ $(r \neq 2)$, by use of B 4.5i. Thus, by B 4.5f, it will suffice to show that no polynomial of degree 3 in the coefficients of a ternary form of order at least 3 (the case r = 1 has effectively been considered in B 4.7b) can vanish whenever the form consists solely of linear factors, unless it vanishes identically. Now let

$\phi(a)$ be a homogeneous cubic polynomial in the coefficients of the ternary form $a_r(x)^r$ and suppose that $\phi(a) = 0$ whenever the form consists solely of linear factors. Then the relation $\phi(a) = 0$ represents a system Φ of curves of order r in a plane with the property that Φ contains all those curves of the plane which consist of r lines. Also, since $\phi(a)$ is of degree 3, Φ has the added property that if it contains four curves of a pencil, then it must contain the whole pencil. It is required to show that if $r \geqslant 3$, then Φ must contain any curve, of order r, in the plane.

Consider, first, the case $r = 3$. Now any non-singular cubic curve in the plane possesses nine inflexions, and these nine inflexions lie on four distinct triangles (see Baker (1) p.20). The triangles each belong to Φ and the original cubic curve belongs to the pencil they determine. Hence the cubic belongs to Φ and the result for $r = 3$ follows. Proceeding by induction on r, suppose that the result has been established for $r = \lambda$ (where $r > 3$). Then it follows that any curve of the plane which consists of a line and a curve of order $r - 1$ must belong to Φ. Now let κ be a generic curve of order r in the plane and let P be a generic point of κ. Let p be a generic line in the plane, meeting κ in $P_1, P_2, P_3, .., P_\lambda$ and let p_1, p_2, p_3 be the respective joins of P to P_1, P_2, P_3. Suppose p_i meets κ again in $P_i', .., P_i^{\lambda-2}$ $(i = 1,2,3)$. Now consider the linear system Ψ of curves of order r which contain all the P's (the P's are $1 + r + 3(r-2) = 4r - 5$ in number). Then the system Ψ is of dimension

at least $\frac{n(n+3)}{2} - (4r - 5) = \frac{n^2-5n+10}{2}$. The curves of this system which contain another point of p and therefore contain p, constitute a linear system of dimension at least $\frac{n^2-5n+8}{2}$ (> 0). Not all these curves contain p_1, because then they would all have to contain p_2 and p_3 also, and the system would have freedom $\frac{(n-4)(n-1)}{2} < \frac{n^2-5n+8}{2}$. Hence there exist curves of Ψ which contain p but not p_1, p_2 or p_3. Similarly, not all curves of Ψ which contain p_1 also contain p_2, for if they did, they would all have to contain p_3 and the system would have freedom $\frac{(n-3)n}{2} - (r - 3) < \frac{n^2-5n+8}{2}$. Hence there are curves of Ψ which contain p_1 but not p_2 or p_3. Similarly there are curves of Ψ containing p_2 but not p_3, and curves of Ψ containing p_3. It follows that if π is a generic curve of Ψ, the pencil determined by π and κ contains four distinct curves, each of which contains one of p, p_1, p_2, p_3. Hence κ belongs to Φ and the result follows.

The proof of B 4.7f is given by B 4.5f and a result due to Brill (2) (Göttingen Nachrichten, Dec. 1893. See also P. Gordan (6), Math. Annalen XLV p.410.) which gives a certain set of polynomials of degree r+1 in the coefficients of an n-ary form of order r $(n > 2)$, whose vanishing is necessary and sufficient for the form to factorize into r linear parts. Brill's formula will be considered again in B4.8e.

The case B 4.7g follows from B 4.7f and B 4.5h.

B 4.7h Remarks. The cases not covered by theorem B 4.7 are $\{\hat{1}\ \hat{\ell}\}$ where $r \geqslant s > 3$, $1 \neq 1$ and $\{\hat{k}\ \hat{\ell}\}$ where $r > 2$ and $k \neq 1 \neq l$. It seems very likely that $\{\hat{1}\ \hat{\ell}\}$ is true whenever $r \geqslant s$. This would follow if the case $\{\hat{1}\ \hat{\lambda-1}\}$ could be shown to be true for each $r = 4, 5, \ldots$ (for, by B 4.5i, $\{\hat{1}\ \hat{\lambda-1}\}$ implies $\{\hat{1}\ \hat{\ell}\}$, and it is evident from B 4.5f that $\{\hat{1}\ \hat{\ell}\}$ implies $\{\hat{1}\ \hat{\ell}\}$ if $r \geqslant s$). By examining the conditions given in B 4.5f or B 4.5e, it is possible to show that $\{\hat{1}\ \hat{\lambda-1}\}$ is equivalent to the statement that a certain numerical determinant (depending on r) does not vanish. The determinant may be obtained as follows :

Consider the different ways of grouping r^2 objects into r groups of r each, the ordering of the groups and of the objects in each group being immaterial. The number of ways of doing this is $\frac{(\lambda^2)!}{(\lambda!)^{\lambda+1}}$. Then each column of the determinant corresponds to one such grouping, and so also does each row. An element of the determinant is defined to be zero if its row and column correspond to groupings of the r^2 objects for which at least one pair of objects is grouped together in both groupings. Otherwise it is defined to be unity.

There is no obvious reason of symmetry why any of these determinants should vanish and, indeed, the determinants given when r = 1, 2, or 3 do not vanish, as is implied by B 4.7b, B 4.7d and B 4.7e. The difficulty in the case r = 4 is evident, as $\frac{(16)!}{(4!)^5} > 2.6 \times 10^6$.

In the cases $\{\hat{1}\ \hat{\ell}\}$ and $\{\hat{j}\ \hat{\ell}\}$, explicit expressions for the Cayley tensors

of the intersection of the two varieties can be obtained by means of B 4.5, from the two formulae :

$$S^{(\lambda,2)} = r2^{\lambda-1}(P^{(\lambda)})^2 \bullet \left\{ \sum_{i=0}^{[\frac{\lambda}{2}]} (-\tfrac{1}{4})^i \frac{(\lambda-i-1)!}{i!(\lambda-2i)!} I^{(2i)} M^{(\lambda-2i,2)} I^{(2i)} \right\} \bullet P^{(2,\lambda)} \bullet (P^{(2)})^\lambda ,$$

$$S^{(3,3)} = \tfrac{1}{2} P^{(3,3)} \bullet (P^{(3)})^3 \bullet \left\{ 72M^{(3,3)} - 162I^{(2)} L^{(2:5)} + 108L^{(1:8)} - 27I^{(5)} L^{(1:3)} + 11I^{(9)} \right\} \bullet P^{(3,3)} \bullet (P^{(3)})^3 .$$

The case $\left\{ \begin{smallmatrix} 2 & 3 \\ 1 & 2 \end{smallmatrix} \right\}$ is the simplest example of a false proposition of type $\left\{ \begin{smallmatrix} \lambda & \mu \\ k & l \end{smallmatrix} \right\}$. (It is false by B 4.7f.) If n = 3, the proposition $\left\{ \begin{smallmatrix} 2 & 3 \\ 1 & 2 \end{smallmatrix} \right\}$ asserts the existence of a polynomial, homogeneous of degree 3 in the coefficients of a ternary quadratic form and homogeneous of degree 2 in the coefficients of the product of three ternary contragredient linear forms, whose vanishing is a necessary and sufficient condition for the conic represented by the quadratic form to contain one of the three points represented by the linear forms. That no such polynomial exists can also be proved directly without much difficulty. Since the varieties $|\mathcal{W}_{f_G}(1,2)|$ and $|\mathcal{W}^{g_G}(2,3)|$ representing, respectively, the conics in a plane and the unordered triads of points in a plane, are both irreducible, this gives a proposition of type $\{f,g\}$ which is false (where f_G is a generic tensor of $\mathcal{W}_{f_G}(1,2)$ and g^{g_G} is a generic tensor of $\mathcal{W}^{g_G}(2,3)$).

B 4.7i It is also possible to obtain from this a false proposition of type $\{f,g\}$ for which the intersection of f and g is not vacuous (i.e. where $k+l > n$). Let f and g be respectively a non-singular quadric and a twisted cubic in a 3-space (n = 4). Now, by B 4.4n, if $\{f,g\}$ were true, there would exist an element V of $\mathcal{F}$ such that for all elements $p \in \mathcal{U}^{g_G}(3)$,

$(f_G)^3 (g_G)^2 \cdot V_\bullet(p)^6 = 0$ is a necessary and sufficient condition for the intersection of the space p with f to meet the intersection of p with g, for all quadrics f and all twisted cubics g. But the intersection of f with p can be any conic in p and similarly the intersection of g with p can be any set of three (non-collinear) points in p. Hence, by B 4.4q and B 4.4r, the truth of $\{f,g\}$ would imply the truth of $\left\{\begin{smallmatrix}1&2&3\\1&2\end{smallmatrix}\right\}$ with $n = 3$, contradicting B 4.7f.

B 4.7j It seems at least possible that the propositions $\{\hat{k}\hat{l}\}$ are all true even when $k \neq 1 \neq l$. If this is the case, then for any two unmixed varieties f and g for which $(f_G, g_G) \notin \mathfrak{G}$, the Cayley form of their intersection can be obtained in either of the following two ways :

Let $f_G \in \mathcal{W}_G(k,r)$, $g_G \in \mathcal{W}_G(1,s)$, $(f_G, g_G) \notin \mathfrak{G}$. Then, by B 4.3, the tensors $(f_G)^{\hat{\ }} \cdot G^{(\hat{k},rs)} (\in \mathcal{W}_G(k,rs))$ and $(g_G)^{\hat{\ }} \cdot G^{(\hat{l},rs)} (\in \mathcal{W}_G(1,rs))$ are Cayley tensors of f counted s times and of g counted r times, respectively. Hence, if $\{\hat{k}\hat{l}\}$ is true, the $(rs)^{\text{th}}$ power of the Cayley form of the intersection of f and g can be obtained by use of B 4.5.

A more satisfactory method, however, is to obtain the required Cayley form as a rational function. Suppose $r \leq s$ and let q be a generic tensor of $\mathcal{U}_G(k)$ with respect to f_G and g_G. Then, by B 4.3, the tensor $f_G(q)^{s-r} \cdot G^{(\hat{k},q)} (\in \mathcal{W}_G(k,rs))$ is a Cayley tensor of the variety which consists of f together with the space q counted s-r times. Now, if $\{\hat{k}\hat{l}\}$ is true, the Cayley form of the intersection of this variety with g can be obtained by use of B 4.5.

Now this Cayley form is the product of the required Cayley form of the intersection of f and g, with $(s-r)^{\text{th}}$ power of the Cayley form of the intersection of the space q with g. But the Cayley form of the intersection of q and g is easily obtained (see B 4.1f). Thus, the required Cayley form, in a frame $\mathscr{F}$, is obtainable as the quotient of two polynomials. The numerator has degree s in the components in $\mathscr{F}$, of each of f_φ and g_φ and degree $s(s-r)$ in the components of q. The denominator has degree s-r in the components of g_φ and $s(s-r)$ in the components of q.

A curious feature of this expression is that although its value is independent of q, it is not always possible to "divide out" the numerator by the denominator to get a polynomial. The reason for this is that the denominator may only be a "factor" of the numerator in virtue of the identical relations satisfied by the elements g_φ of $\mathscr{W}_\varphi(k,r)$ (see B 4.2). That such a situation actually arises can be seen by considering the case $k = 1$, $l = 2$, $r = 2$, $s = 3$, since by B 4.7, $\left\{\begin{smallmatrix}3&3\\1&2\end{smallmatrix}\right\}$ is true whereas $\left\{\begin{smallmatrix}2&3\\1&2\end{smallmatrix}\right\}$ is false.

It is perhaps worth examining more closely what happens in this particular case. Suppose for simplicity that $n = 3$, and that $f_\varphi \in \mathscr{W}_\varphi(1,2)$, $g^\varphi \in \mathscr{W}^\varphi(2,3)$, $q \in \mathscr{U}_\varphi(1)$ $(=\zeta_\varphi)$. Then, by B 4.5a, the condition for f to meet g is $\phi(f_\varphi, g^\varphi) = 0$ where

$$\phi(f_\varphi, g^\varphi) = \frac{(f_\varphi q \cdot p^{(3)})^3 \cdot S^{(3,3)} \cdot (g^\varphi)^3}{(q)^3 \cdot g^\varphi}.$$

Now it can be verified that if $h^{\alpha 3}$ is any symmetric element of $\mathcal{C}_{\alpha\beta 3}$, then

$$(f_\phi q \cdot P^{(3)})^3 \cdot S^{(3,3)}_{\cdot\cdot}(h)^3 = \{(q)^3 \cdot h\}\{(f_\phi)^3 \cdot S^{(3,3)}_{\cdot\cdot}(h)^2\} - 36\{(f_\phi)^3 \cdot M^{(3,3)}_{\cdot\cdot}(\epsilon)^2\}\{(\epsilon^*)^2 (q)^3 \cdot M^{(3,3)}_{\cdot\cdot}(h)^3\}$$

But $(q)^3 \cdot h$ is not always a factor of the right hand side of this because it is not always a factor of $(\epsilon^*)^2 (q)^3 \cdot M^{(3,3)}_{\cdot\cdot}(h)^3$. However, it will be observed that $(\epsilon^*)^2 (x)^3 \cdot M^{(3,3)}_{\cdot\cdot}(h)^3$ (where $x_\alpha \epsilon \mathcal{C}_\alpha$) is the Hessian of the cubic form $(x)^3 \cdot h$, and that if this cubic form is the product of three linear factors, then it is proportional to its Hessian. Now this will be so if $h = g^\phi$ because $(x)^3 \cdot g^\phi$ is the Cayley form of g and is thus the product of the Cayley forms of the three points which constitute g. It follows that the expression $\frac{(\epsilon^*)^2 (q)^3 \cdot M^{(3,3)}_{\cdot\cdot}(g^\phi)^3}{(q)^3 \cdot g^\phi}$ is independent of q. It cannot be a simple invariant of g^ϕ, however, because a ternary cubic possesses no invariant of degree 2. It is an example of what might be defined as a <u>rational invariant</u> of a tensor.

The expression $\phi(f_\phi, g^\phi)$ is thus also independent of q, and it is a rational invariant of f_ϕ and g^ϕ. Further, since $\left\{\begin{smallmatrix} 4 & 3 \\ 1 & 2 \end{smallmatrix}\right\}$ is true by B 4.7e, we have

$$\{\phi(f_\phi, g^\phi)\}^2 = ((f_\phi)^2 \cdot P^{(3)})^3 \cdot S^{(4,3)}_{\cdot\cdot}(g^\phi)^4 .$$

This shows that the variety $\mathcal{X}\left(\begin{smallmatrix} 2 & 3 \\ 1 & 2 \end{smallmatrix}\right)$, though not a complete intersection of $\mathcal{W}_\phi(1,2) \times \mathcal{W}^\phi(2,3)$ with a primal, has the property that, counted with multiplicity 2, it is such an intersection. Also, by the preceding arguments, we see that it is possible to add to $\mathcal{X}\left(\begin{smallmatrix} 2 & 3 \\ 1 & 2 \end{smallmatrix}\right)$ a complete intersection so that the resulting variety is also a complete intersection.

B 4.8 Theorem. Let $0 \neq a^{\alpha_{\lambda_1}} \in \mathcal{C}_{\alpha \times \lambda_1}$ Then a necessary and sufficient condition for a to belong to $\mathcal{W}^H(n-s,r)$ is that $a = H^{(\lambda_1,\lambda_1)}_{\bullet} a$ and

$$K^{(\alpha+\lambda_1,\lambda_1)} P^{(\alpha(\lambda_1-1))} (P^{(\lambda_1^2)})^{\lambda_1-1} \bullet I^{(\lambda_1 \alpha)} S^{(\alpha,\lambda_1)} I^{(\lambda_1^2(\alpha-1))} \bullet I^{(\alpha\lambda_1)} M^{(\alpha,\lambda_1,\lambda_1)}_{\bullet}(a)^{\lambda_1+1} = 0.$$

Proof. By B 3.5 i the condition for a to be a tensor of type a^H corresponding to a complex of (n-s)-spaces of order r in $\mathfrak{P}^s$ is $a = H^{(\alpha,\lambda_1)}_{\bullet} a$. It is required to show that the condition for this complex to be a set of (n-s)-spaces which meet an unmixed variety in $\mathfrak{P}^s$ is that a satisfy the other relation given in B 4.8. (The variety must then necessarily have dimension s-1 and multiplicative order r, as required.) This complex Λ is given by $(u_{(1)})^{\lambda} .. (u_{(\alpha)})^{\lambda} \bullet a = 0$ where $u_{(1)}, .., u_{(\alpha)}$ represent s primes through an (n-s)-space of Λ. Now the condition on for it to be the set of (n-s)-spaces meeting such a variety, is that for each $u_{(1)} \bullet , , \bullet u_{(\alpha)}$, the form $(u)^{\lambda}(u_{(2)})^{\lambda} .. (u_{(\alpha)})^{\lambda} \bullet a$ factorizes into linear factors

$$(u)^{\lambda}(u_{(2)})^{\lambda} .. (u_{(\alpha)})^{\lambda} \bullet a = (u \bullet b_{(1)}) .. (u \bullet b_{(\lambda_1)})$$

$(b^{\alpha}_{(1)}, .., b^{\alpha}_{(\lambda_1)} \in \mathcal{C}$ and they depend on $u_{(2)}, .., u_{(\alpha)})$ and that for each factor $u \bullet b_{(i)}$, the point $b_{(i)}$ is singular for Λ, i.e. by B3.5 r,

$$K^{(\alpha+1,\lambda_1)}_{\bullet} a (b_{(i)})^{\lambda} = 0.$$

(See (10) pp.51-57.)

Suppose first, that Λ has this property (ie $a \in \mathcal{W}^H(n-s,r)$). Then for each $u_{(1)}, .., u_{(\alpha)}$ we have, for suitable $b_{(1)}, .., b_{(\lambda_1)}$

$$P^{(\lambda)} \cdot b_{(1)} \ldots b_{(\lambda)} = P^{(\lambda)} \cdot (u_{(1)})^{\lambda} \ldots (u_{(\delta)})^{\lambda} \cdot a$$
$$= I^{(\lambda)} (u_{(1)})^{\lambda} \ldots (u_{(\delta)})^{\lambda} \cdot a$$

by B 1.2 and the symmetry of a. Hence

$$\left\{ S^{(\lambda,\lambda)} (u_{(1)})^{\lambda^2} \ldots (u_{(\delta)})^{\lambda^2} \right\} \cdot M^{(\lambda;\lambda,\delta)} \cdot (a)^{\lambda} = S^{(\lambda,\lambda)} \cdot \left\{ I^{(\lambda)} (u_{(1)})^{\lambda} \ldots (u_{(\delta)})^{\lambda} \right\}^{\lambda} \cdot (a)^{\lambda}$$
$$= S^{(\lambda,\lambda)} \cdot \left\{ P^{(\lambda)} \cdot b_{(1)} \ldots b_{(\lambda)} \right\}^{\lambda}$$

so that, by B 4.6b,

$$\left\{ P^{(\lambda)} P^{(\lambda(\lambda-1))} (u_{(1)})^{\lambda^2} \ldots (u_{(\delta)})^{\lambda^2} \right\} \cdot S^{(\lambda,\lambda)} I^{(\lambda^2(\delta-1))} \cdot M^{(\lambda;\lambda,\delta)} \cdot (a)^{\lambda} = \sum_{i=1}^{\hat{r}} (b_{(i)})^{\lambda} c_{(i)}$$

where $c_{(i)} \in \mathcal{C}_{u^{*\lambda(\delta-1)}}$. But $K^{(\delta+1,\lambda)} \cdot a (b_{(i)})^{\lambda} = 0$, $(i = 1, \ldots, r)$, whence

$$K^{(\delta+1,\lambda)} I^{(\lambda(\delta-1))} \cdot \left\{ a\ P^{(\lambda)} P^{(\lambda(\delta-1))} (u_{(1)})^{\lambda^2} \ldots (u_{(\delta)})^{\lambda^2} \right\} \cdot S^{(\delta,\lambda)} I^{(\lambda^2(\delta-1))} \cdot M^{(\lambda;\lambda,\delta)} \cdot (a)^{\lambda} = 0.$$

This must hold for each $u_{(1)}, \ldots, u_{(\delta)} \in \mathcal{C}_u$, so that, by B 1.2 and the symmetry of $K^{(\delta+1,\lambda)}$ we get

$$K^{(\delta+1,\lambda)}\ P^{(\lambda(\delta-1))}\ (P^{(\lambda^2)})^{\delta-1} \cdot I^{(\delta\lambda)} S^{(\lambda,\delta)} I^{(\lambda^2(\delta-1))} \cdot I^{(\delta\lambda)} M^{(\lambda;\lambda,\delta)} \cdot (a)^{\lambda^2+1} = 0$$

as required.

Suppose, conversely, that a satisfies this relation (and $a = H^{(\lambda,\delta)} \cdot a$). Then, for each $u_{(1)\alpha}, \ldots, u_{(\delta)\alpha}$, we have

$$K^{(\delta+1,\lambda)} I^{(\lambda(\delta-1))} \cdot \left\{ a\ P^{(\lambda)} P^{(\lambda(\delta-1))} \right\} \cdot S^{(\lambda,\lambda)} \cdot \left\{ I^{(\lambda)} (u_{(1)})^{\lambda} \ldots (u_{(\delta)})^{\lambda} \cdot (a)^{\lambda} = 0, \right.$$

that is, $$K^{(\delta+1,\lambda)}\ I^{(\lambda(\delta-1))} \cdot ad = 0 \qquad \text{B 4.8a}$$

where $$d = P^{(\lambda)} P^{(\lambda(\delta-1))} \cdot S^{(\lambda,\lambda)} \cdot (f)^{\lambda}, \qquad \text{B 4.8b}$$

$$f = \left\{ I^{(\lambda)} (u_{(1)})^{\lambda} \ldots (u_{(\delta)})^{\lambda} \right\} \cdot a.$$

Now the primes u ($u_{\alpha} \in \mathcal{C}_u$) which satisfy $(u)^{\lambda} \cdot f = 0$ are just tjose whose intersection with $u_{(1)}, \ldots, u_{(\delta)}$ is an (n-s)-space of Λ, provided that $u, u_{(1)}, \ldots, u_{(\delta)}$ are linearly independent. Suppose that the primes $u_{(1)}, \ldots, u_{(\delta)}$

are linearly independent and that their (n-s+1)-dimensional join p is such that it contains an (n-s)-space not belongigg to Λ (otherwise f = 0). Then $(u)^{\lambda} \cdot f = 0$ is the condition that the prime u should contain an (n-s)-space of Λ, which lies in p. Now the (n-s)-spaces of Λ in p constitute a complex of primes in p. This complex is a (multiplicative) envelope of primes which touch a variety Γ, where Γ is a variety in p having no linear components except possibly for a set of points of p. Then Γ is also similarly enveloped by the set of primes u in $\mathfrak{P}$ for which $(u)^{\lambda} \cdot f = 0$. (If s = 0, p = $\mathfrak{P}$.) It is required to show that consists only of a set of points (for then $f = P^{(\lambda)} \cdot b_{(1)} \ldots b_{(\lambda)}$ where Γ consists of the points $b_{(1)}, \ldots, b_{(\lambda)}$), each of which is singular for

Let Γ' be an irreducible component of Γ and let the multiplicity of the system of primes touching Γ', in the envelope $(u)^{\lambda} \cdot f = 0$, be m. Let v be a generic prime of the envelope touching Γ', and let x be contact point of v with Γ' (v_α, $x^\alpha \epsilon \mathcal{C}$). Now if m = 1, the point x is given by

$$x = I^{(1)} (v)^{\lambda - 1} \cdot f$$

(since $I^{(1)}(v)^{\lambda-1} \cdot f = \frac{1}{\lambda}\left[\frac{\partial}{\partial u_\alpha}\left\{(u)^{\lambda} \cdot f\right\}\right]_{u=v}$) and we have $(v)^{\lambda} \cdot f = 0$. Suppose $m \geqslant 1$ and $(u)^{\lambda} \cdot f \equiv \left\{(u)^{\lambda'} \cdot f'\right\}^{m} \left\{(u)^{\lambda''} \cdot f''\right\}$, f' being the symmetric tensor corresponding to Γ'. We have $(v)^{\lambda'} \cdot f' = 0$, $(v)^{\lambda''} \cdot f'' \neq 0$ and $x = I^{(1)} (v)^{\lambda'-1} \cdot f'$, so that

$$I^{(m)}(v)^{\lambda - m} \cdot f = I^{(m)}(v)^{\lambda - m} \cdot P^{(\lambda)} \cdot (f')^{m} f'' = \frac{(\lambda')^{m}}{(\lambda)_m} \left\{I^{(1)} (v)^{\lambda'-1} \cdot f'\right\}^{m} \left\{(v)^{\lambda''} \cdot f''\right\}$$

since, when $P^{(\lambda)}$ is expanded, the other terms give zero. Hence, for some $\xi \neq 0$,

CHAPTER 6

$$(x)^m = \xi\, I^{(m)}(v)^{n-m} \cdot f.$$

Also,

$$I^{(m-1)}(v)^{n-m+1} \cdot f = 0$$

It follows that

$$I^{(nm)}(v)^{n^2-nm} \cdot \{P^{(n)} P^{(n^2-n)} \cdot S^{(n,n)}\}(f)^n = \sum_{i=1}^{(n^2)!} \xi_i I^{(nm)}(v)^{n^2-nm} \cdot J_i \cdot (f)^n = \xi \{I^{(m)}(v)^{n-m} \cdot f\}^n$$

since the other terms vanish, $(\xi, \xi_1, \xi_2, \ldots \in \mathfrak{R})$. That is, by B 4.8b,

$$I^{(nm)}(v)^{n^2-nm} \cdot d = \eta (x)^{nm} \qquad \text{B 4.8c}$$

where $\eta = \xi \xi^n$

I shall next show that $I^{(nm)}(v)^{n^2-nm} \cdot d \neq 0$. It will be sufficient to do this in the case when the form $(u)^n \cdot f$ consists of linear factors. For it is well-known that any algebraic variety can be specialized to a set of linear spaces, whose multiplicities are equal to those of the corresponding components of the variety (see (10) pp.141-145).

Suppose $(u)^n \cdot f \equiv (u \cdot g_{(1)}) \ldots (u \cdot g_{(n)})$ where $g_{(1)}^{\alpha}, \ldots, g_{(n)}^{\beta} \in \mathfrak{G}_{\mathfrak{a}^*}$, that is,

$$f = P^{(n)} \cdot g_{(1)} \ldots g_{(n)}.$$

Let $g_{(1)} = \ldots = g_{(m)} \neq g_{(i)}$ for $i = m+1, \ldots, r$ and let $v \cdot g_{(1)} = 0$, $v \cdot g_{(i)} \neq 0$ $(i = m+1, \ldots, r)$. (That is, f' is specialized to $P^{(n)} \cdot (g_{(1)} g_{(m+1)} \ldots)$, say, and v to a prime through the point $g_{(1)}$ but not through any of the other g's.) Then, by B 4.6b,

$$\begin{aligned} I^{(nm)}(v)^{n^2-nm} \cdot d &= I^{(nm)}(v)^{n^2-nm} \cdot \{P^{(n)} P^{(n^2-n)} \cdot S^{(n,n)}\} \cdot (P^{(n)} \cdot g_{(1)} \ldots g_{(n)})^n \\ &= \frac{1}{n} \sum_{i=1}^{n} (g_{(i)})^n \{I^{(nm-n)}(v)^{n^2-nm} \cdot P^{(n^2-n)} \cdot (g_{(1)})^n \ldots (g_{(i-1)})^n (g_{(i+1)})^n \ldots (g_{(n)})^n \\ &= \frac{m}{n\binom{n^2-n}{nm-n}} (g_{(1)})^{nm} (v \cdot g_{(m+1)})^n \ldots (v \cdot g_{(n)})^n \neq 0 \end{aligned}$$

as required, so that the η of B 4.8c is non-zero.

From B 4.8a and B 4.8c we get

$$\eta \{K^{(a+1,\lambda)} \cdot a(x)^{\lambda}\}(x)^{\lambda m-\lambda} = K^{(a+1,\lambda)} I^{(\lambda m)} (v)^{\lambda^2-\lambda m} \cdot ad = 0,$$

whence $K^{(a+1,\lambda)} \cdot a(x)^{\lambda} = 0$, that is x is singular for Λ (by B 3.5r). Thus every point of Γ is singular for Λ. But if Γ were not purely zero-dimensional, this would imply that every (n-s)-space in p belong to Λ, contrary to hypothesis. This shows that Γ consists only of a set of points, each of which is singular for Λ, and proves the result.

B 4.8d Corollary. Let a_{λ}, $x^a \epsilon \mathcal{C}$, $a = a \cdot P^{(\lambda)}$. Then a necessary and sufficient condition for the form $a \cdot (x)^{\lambda}$ to factorize completely into linear factors is

$$(a)^{\lambda+1} \cdot I^{(\lambda)} S^{(\lambda,\lambda)} \cdot K^{(s,\lambda)*} P^{(\lambda^2-\lambda)} = 0.$$

Proof. The condition for $a \cdot (x)^{\lambda}$ to split into linear factors is that $a^* \epsilon \mathcal{W}^{n}(n-1,r)$, since the Cayley forms of zero-dimensional varieties are just those forms $(u)^{\lambda} \cdot a^*$ which split into linear factors. Hence the result follows from B 4.8 with s = 1.

B 4.8e Remark. The formula given in B 4.8d is essentially Brill's formula(2) in a different notation.

Brill states his condition in the symbolic form $(apuv..)^{\lambda} = 0$ where $(a_x)^{\lambda}$ is the given form and where $(p_x)^{\lambda}$ is a certain covariant of degree r in this form, in two sets of variables x and y ; $(p_x)^{\lambda} = g(x,y)$ $= r^{\lambda-1} f_x^{\lambda} - r^{\lambda-1} \frac{\lambda-1}{2} f f_x^{\lambda-2} f_{x^2} + \ldots$ where $f = f(y) = (a_y)^{\lambda}$. The expression

$g(x,y)$ has the defining property that if the roots of $f(x - \lambda y)$ are $\lambda_1,..,\lambda_\lambda$, then $g(x,y) = \frac{1}{\lambda}(\lambda_1^{\wedge} + .. + \lambda_\lambda^{\wedge})f^{\wedge}$. It may be verified, using B 4.6b, that $g(x,y) = (a)^{\lambda} \bullet S^{(\lambda,\mu)} \bullet (x)^{\lambda} (y)^{\mu-\lambda}$ and hence that Brill's formula is equivalent to that given in B 4.8d.

It is also possible to give a proof of B 4.8 from Brill's formula. Brill's result is, in fact, the essential part of B 4.8.

References

(1) Baker, H. F. Principles of Geometry, Vol.5 (Cambridge) : 1933.

(2) Brill, Göttingen Nachrichten : December 1893.

(3) Brown B. & McCoy, N. H. The maximal regular ideal of a ring. Proc. Amer. Math. Soc. 1, 165-171: 1950.

(4) Dwyer, P. S. Linear Computations, PP.225-8. (New York) : 1951.

(5) Eisenhart, L. P. Riemannian Geometry, (Priceton) : 1926.

(6) Gordan, P. Math Annalen 45, p410 : 1894.

(7) Hodge, W. V. D. Harmonic Integrals, pp.17-2[illegible]. (Cambridge) : 1952

(8) Hodge, W. V. D. A note on k-connexes. Proc. Camb. Phil. Soc. 38, (1942), 129.

(9) Hodge, W. V. D. & Pedoe, D. Methods of Algebraic Geometry, vol.1 (Cambridge) : 1947

(10) Hodge, W.V. D. & Pedoe, D. Ibid., vol2 (Cambridge) : 1952.

(11) Littlewood, D. E. The Theory of Group Characters and Matrix Representations of Groups, 2nd Ed. (Oxford) : 1950.

(12) Littlewood, D. E. Invariant theory , tensors and group characters. Phil. Trans. A, 239, 305.

(13) Moore, E. H. General analysis Pt 1. Mem. Amer Phil.Soc. pp.8, 197-209: 1935.

(14) Moore, E. H. Bull. Amer Math. Soc. (2) 26, 394-5 : 1920.

(15) von Neumann, J. On regular rings. Proc. Nat. Acad. Sci. 22, 707-713 (1936)

(16) Pedoe, D. On a new analytical representation of curves in space. Proc. Camb. Phil. Soc. 43, 455-8 : (1947).

(17) Penrose, R. A generalized inverse for matrices. Proc. Camb. Phil. Soc. 51, 406-13 : (1955).

(18) Penrose, R. On best approximate solutions of linear matrix equations. Proc. Camb. Phil. Soc. 52, 17-19 : 1956.

(19) Rado, R. Note on generalized inverses of matrices. Proc. Camb. Phil. Soc. 52, 600-601 : 1956.

(20) Rutherford, D. E. Substitutional Analysis. (Edinburgh)1948.

(21) Turnbull, H. W. The Theory of Determinants Matrices and Invariants, 2nd Ed. (Glasgow) : 1945.

(22) van der Waerden, B. L. Modern Algebra, vol.1 (New York) : 1949

(23) van der Waerden, B. L. Ibid., vol.2. New York) : 1950.

(24) van der Waerden, B. L. Zur Konstruction des Resultantensystems für homogene Gleichungen. Arch. Math. 5, 371-375 (1954).

7
GEOMETRICAL ALGEBRAS: A NEW APPROACH TO INVARIANT THEORY

Fellowship Thesis, St John's College, Cambridge 1957

The Cayley-form problem of Chapter 6 raised the more general issue of how to construct a tensor quantity that is determined in a geometrical or invariant way from a given finite set of tensor objects. The standard theory of invariants—or of tensor operations—provides the correct answer only in certain situations, such as when there are no algebraic constraints on the given set. This Fellowship thesis provides a neat answer to this class of problem, in terms of quotients or factors of tensorial expressions. In the course of this work I was led to a notion of a 'generic' object, relative to a given set of objects. I never followed up the issue of possible connections with the ideas of Grothendieck. Nor did I publish the results of either thesis.

Preface

The approach described here of basing an invariant theory entirely on the formal algebraic properties of tensors and of tensor rays is, as far as I know, original. The same applies to the theory given here of geometrical objects and geometrical functions. Most of the results of the rest of the dissertation are based on these ideas and are therefore also original. Apart from some results given in my Ph.D. dissertation, references are given where known results, other than elementary ones, are used or proved.

I should like to thank Dr J. A. Todd for advice and encouragement given me while the greater part of this research was carried out.

January 1957

Bedford College,
London.

Geometrical algebras : a new approach to invariant theory.

by R.Penrose

1. Introduction. One of the aims of the classical theory of invariants is to isolate those algebraic properties of a set of forms, which are geometrical in nature. For this purpose the symbolic calculus has been developed (see, for example, (2)). An alternative, but related approach is the method of tensors (see (6)). The interests of workers in this field have been largely concerned with questions relating to the complete system of concomitants for a set of general forms. However, many of the forms which arise from geometrical concepts (e.g Cayley forms) are not general forms but satisfy identical algebraic relations. In these circumstances it appears that the classical approach is, to a certain extent, inadequate, since it turns out (see 6.11, 8.6) that, in many cases the concomitants of a set of forms do not include all of the forms which arise from the original forms in a geometrical manner nor even all those which are left absolutely invariant by the linear transformations which leave the original forms invariant. (Also, this situation has a tendency to become obscured if all forms considered are reduced to irreducible parts.)

The approach given here is based on a systematic application of the algebraic properties of tensors, and of sets of tensors which are closed

under the tensor operations. In order to facilitate the handling of these tensors a notation is introduced which enables groups of tensor indices to be regarded as single indices and also enables indices to be dispensed with in many cases. Thus, the generality of results is greatly increased without added notational inconvenience. The tensors are to be regarded essentially as abstract entities obeying formal rules of manipulation. The components of a tensor are then just certain scalars which arise as (tensor) functions of the given tensor and the vectors giving the frame of reference (see 4.2).

In order to get a more exact correspondence between the geometry and the algebra, the concept of a tensor ray system is introduced. This is an algebraic structure which, to a certain extent, replaces the field as the basic algebraic system of the geometry. Each geometrical object is associated with a single element of the tensor ray system, the geometrical properties of these objects and the algebraic properties of the tensor rays exactly corresponding to one another. Some criteria are obtained for determining algebraically those objects definable geometrically from a given set of geometrical objects. Also, in order that geometries other than projective geometry may be treated, a corresponding treatment is given for the related problem of determining the tensors left invariant by the group of linear transformations which leaves a given set of tensors invariant. The results described here enable an answer to be

given to a conjecture due to D. E. Littlewood (7) concerning this problem.

2. The tensor operations.

A set of numbers (or functions) $a^{i_1 \cdots i_r}_{j_1 \cdots j_s}$ ($i_1, .., i_r$, $j_1, .., j_s$ run from $1, .., n$) is said to constitute the set of components of a tensor of orders (s,r) (or rank $r+s$) if they transform according to certain transformation laws (see (1) p 8a). The four basic operations which can be performed on tensors to produce new tensors are those of addition, outer multiplication, contraction and permutation of indices. The set of tensors which can be obtained from a given set of tensors by application of the tensor operations is closed with respect to these operations and thus constitutes an algebraic system admitting these four operations. I shall be concerned here with the algebraic properties of these tensor operations and thus, a tensor will be regarded simply as an algebraic entity subject to certain formal rules of manipulation.

2.1 Notation. A symbol a, without indices, may be used to denote the tensor whose components (in some fixed reference frame) are $a^{i_1 \cdots i_r}_{j_1 \cdots j_s}$. I shall also use a symbol $a^{\alpha_1 \cdots \alpha_r}_{\alpha_1 \cdots \alpha_s}$ to denote this same tensor. The indices $\alpha_1, .., \alpha_r$ and $\alpha_1, .., \alpha_s$ are simply markers and are not supposed to be symbols standing for $1, .., n$. The upper α's and lower α's must each be numbered with integers in their natural order if the symbol is to represent the tensor a. (With this notation repeated Greek indices do not indicate contraction, see 2.4.)

CHAPTER 7

2.2 Transposition. If a is as in 2.1 and the tensor b is given by

$$b^{i_1 \cdots i_r}_{j_1 \cdots j_s} = a^{i_{k(1)} \cdots i_{k(r)}}_{j_{l(1)} \cdots j_{l(s)}}$$

for each $i_1,..,i_r$, $j_1,..,j_s$, where k and l denote permutations of 1,..,r and of 1,..,s respectively then b will be called a (generalized) transpose of a. I shall write

$$b = b^{\alpha_1 \cdots \alpha_r}_{\alpha_1 \cdots \alpha_s} = a^{\alpha_{k(1)} \cdots \alpha_{k(r)}}_{\alpha_{l(1)} \cdots \alpha_{l(s)}}$$

(Thus, $a^{\alpha_2 \alpha_1 \alpha_3 \cdots \alpha_r}_{\alpha_1 \cdots \alpha_s}$ and $a^{\alpha_1 \cdots \alpha_r}_{\alpha_3 \alpha_2 \alpha_1 \alpha_4 \cdots \alpha_s}$ are examples of transposes of a.) This operation is a generalization of the idea of forming the transpose of a matrix. It is because of this operation that it is necessary to number the indices of a with consecutive integers. For if the tensor $c_{\alpha_1 \alpha_2}$ were denoted by $c_{\beta\gamma}$, say, it would lead to confusion to denote the transpose $c_{\alpha_2 \alpha_1}$ of c by $c_{\gamma\beta}$ (but see 2.8). The general transpose of a tensor is built up by means of elementary transposition of pairs of upper or pairs of lower indices.

A set of such indices may be regarded as an operator acting on the tensor. It is the identity operator if the indices are numbered in their natural order.

2.3 Multiplication. If a and b are given by the sets of components $a^{i_1 \cdots i_r}_{j_1 \cdots j_s}$ and $b^{p_1 \cdots p_t}_{q_1 \cdots q_u}$, respectively, the tensor c whose components are given by

$$c^{i_1 \cdots i_r p_1 \cdots p_t}_{j_1 \cdots j_s q_1 \cdots q_u} = a^{i_1 \cdots i_r}_{j_1 \cdots j_s} \, b^{p_1 \cdots p_t}_{q_1 \cdots q_u}$$

for each $i_1,..,q_u$, is the (outer) product of a with b. I shall denote this by

$$c = ab = c^{\alpha_1 \cdots \alpha_{r+t}}_{\alpha_1 \cdots \alpha_{s+u}} = a^{\alpha_1 \cdots \alpha_r}_{\alpha_1 \cdots \alpha_s} \, b^{\alpha_{r+1} \cdots \alpha_{r+t}}_{\alpha_{s+1} \cdots \alpha_{s+t}}$$

Multiplication is clearly associative but not commutative and there are

no divisors of zeros. (The zeros are, of course, the tensors all of whose components are zero.)

This definition of multiplication will have to be modified slightly in 2.9, however.

2.4 Contraction. If the tensor a is given by the set of components $a^{i_1 \cdots i_\lambda}_{j_1 \cdots j_\mu}$, new tensors may be formed by "contracting" upper indices with lower indices. Thus, if the components of c are given by

$$c^{i_1 \cdots i_{\lambda-1}}_{j_1 \cdots j_{\mu-1}} = \sum_{m=1}^{n} a^{i_1 \cdots i_{\nu-1}\, m\, i_\nu \cdots i_{\lambda-1}}_{j_1 \cdots j_{\nu-1}\, m\, j_\nu \cdots j_{\mu-1}}$$

for each $i_1, \ldots, j_{\mu-1}$, then c is also a tensor : a contraction of a. (The summation sign is usually omitted.) I shall denote this contraction by

$$c = c^{\alpha_1 \cdots \alpha_{\lambda-1}}_{\alpha_1 \cdots \alpha_{\mu-1}} = a^{\alpha_1 \cdots \alpha_{\nu-1}\, \bar{\alpha}_1\, \alpha_\nu \cdots \alpha_{\lambda-1}}_{\alpha_1 \cdots \alpha_{\nu-1}\, \bar{\alpha}_1\, \alpha_\nu \cdots \alpha_{\mu-1}} .$$

The general transpose of c may then be written

$$c^{\alpha_{k(1)} \cdots \alpha_{k(\lambda-1)}}_{\alpha_{l(1)} \cdots \alpha_{l(\mu-1)}} = a^{\alpha_{k(1)} \cdots \alpha_{k(\nu-1)}\, \bar{\alpha}_1\, \alpha_{k(\nu)} \cdots \alpha_{k(\lambda-1)}}_{\alpha_{l(1)} \cdots \alpha_{l(\nu-1)}\, \bar{\alpha}_1\, \alpha_{l(\nu)} \cdots \alpha_{l(\mu-1)}} .$$

(Barred indices are "dummy indices".) A second contraction of a may be written

$$a^{\alpha_1 \cdots \alpha_{\nu-1}\, \bar{\alpha}_1\, \alpha_\nu \cdots \alpha_{\nu'-1}\, \bar{\alpha}_2\, \alpha_{\nu'} \cdots \alpha_{\lambda-2}}_{\alpha_1 \cdots \alpha_{\nu'-1}\, \bar{\alpha}_2\, \alpha_{\nu'} \cdots \alpha_{\nu-1}\, \bar{\alpha}_1\, \alpha_\nu \cdots \alpha_{\mu-2}} .$$

The symbol $\bar{\alpha}_3$ may be used for a third contraction and so on. We have

$$a^{\cdots \bar{\alpha}_1 \cdots \bar{\alpha}_2 \cdots \bar{\alpha}_3 \cdots}_{\cdots \bar{\alpha}_2 \cdots \bar{\alpha}_1 \cdots \bar{\alpha}_3 \cdots} = a^{\cdots \bar{\alpha}_3 \cdots \bar{\alpha}_2 \cdots \bar{\alpha}_1 \cdots}_{\cdots \bar{\alpha}_2 \cdots \bar{\alpha}_3 \cdots \bar{\alpha}_1 \cdots}$$

etc. The general transpose of a contracted tensor is obtained by permuting the unbarred α's.

2.5 Addition. Tensor addition (and subtraction) completes the list of tensor operations. Two tensors are conformable for addition if they have the same number of uncontracted covariant and of uncontracted contravariant indices. The sum of two tensors a and b is denoted by a+b and has as components the sums of the corresponding components of a and of b.

Tensors form a kind of graded ring with respect to addition and multiplication – but addition between tensors of different types is not permitted.

2.6 The semigroups of index types. I shall call the lower (covariant) indices of a tensor $a^{\alpha_1 \cdots \alpha_r}_{\alpha_1 \cdots \alpha_s}$ indices of type α and the upper (contravariant) ones, indices of type α^*. It will prove very convenient to be able to introduce new index types in addition to α and α^*, which are defined formally as the (commutative) product of α's and α^*'s. Thus, some of the indices of a tensor may be grouped together in the form of a composite index whose type is the product of the index types of the indices (covariant and contravariant) involved. The index types thus form a free commutative semigroup, which I shall denote by $\mathfrak{G}$, the generators of this semigroup being α and α^*. I shall suppose that $\mathfrak{G}$ also possesses an identity 1, which is not used as an index. The general element β of $\mathfrak{G}$ is of the form $\beta = \alpha^p \alpha^{*q}$ $(p, q \geqslant 0)$. The element $\alpha^q \alpha^{*p}$ will be called the conjugate of β and will be written β^*. We have $(\beta^*)^* = \beta$ and $(\beta\gamma)^* = \beta^* \gamma^*$

Any element β of $\mathfrak{G}$ may be used as an index of a tensor. Thus if $\gamma = \alpha\alpha^*$ and $\delta = \alpha^3$, the symbol $b_{\gamma\delta}{}^{\beta\gamma\delta\beta^*}$ stands for a tensor with two covariant indices and five contravariant indices. A composite index β may also be

used as an upper index. An upper index β is of type β^*. Thus the symbol $c_{\gamma}^{\delta *}$ stands for a tensor with four covariant indices and one contravariant index (see also 2.8).

Any tensor is associated with an element of $\mathfrak{G}$, namely that element β which is the product of all the index types of indices possessed by the tensor. I shall call β the <u>type</u> of the tensor. A tensor $a_{\alpha_1 \cdots \alpha_s}^{\alpha_1 \cdots \alpha_r}$ of orders (r,s) is thus of type $\alpha^r \alpha^{*s}$. The word rank will, however, be required for tensors in its matrix sense. The scalars are the tensors of type 1 (the identity of $\mathfrak{G}$).

2.7 <u>Notation</u>. For notational reasons (see 2.10) it will be found convenient to associate a symbol a, b, or c, etc., representing a tensor, not only with that tensor but also with a particular division of its indices into two groups, upper and lower. (In the case of a scalar, however, this consideration is trivial.) If it is required to consider two different groupings of indices for the same tensor, two different symbols must be used. Thus, for example, the same tensor might be denoted by a, b and c, where

$$a_{\gamma}^{\beta} = b_{\beta^* \gamma} = c^{\beta \gamma^*}$$

2.8 <u>Composite indices</u>. The Greek letters $\beta, \gamma, \ldots$ denoting elements of $\mathfrak{G}$ may be given numerical suffixes when used for indices of tensors, like the letter α. The same applies to $\alpha^*, \beta^*, \gamma^*, \ldots$ If letters $\alpha, \alpha^*, \beta, \beta^*, \gamma, \ldots$ are used as lower indices attached to a tensor symbol a, with

no upper indices, and if the numerical suffixes of each of $\alpha, \alpha^*, \beta, \beta^*, \gamma, \ldots$ appear in their natural order, then this symbol represents the tensor a.

For example, if $\beta = \alpha\alpha^*$, $\gamma = \alpha^2$, and a is of type $\alpha\alpha^*\beta^2\gamma^*(= \alpha^3\alpha^{*5})$, then

$$a = a_{\alpha_1\alpha_1^*\beta_1\beta_2\gamma_1^*} = a_{\beta_1\gamma_1^*\alpha_1^*\beta_2\alpha_1}$$
$$= a_{\alpha_1\alpha_1^*\alpha_2\alpha_2^*\alpha_3\alpha_3^*\alpha_4^*\alpha_5^*} = a_{\alpha_1\alpha_2\alpha_3\alpha_1^*\ldots\alpha_5^*}$$

The order to which the covariant and contravariant indices appear relative to each other is not relevant. In general, each symbol $\beta_i, \ldots$ stands for a set of covariant and contravariant indices. If the β's are permuted among themselves, then the symbol obtained stands for that transpose of a which is obtained by performing the corresponding permutation on the covariant and contravariant indices concerned. Thus, if a is as above, we have

$$a_{\alpha_1\alpha_1^*\beta_2\beta_1\gamma_1^*} = a_{\alpha_1\alpha_1^*\alpha_3\alpha_3^*\alpha_2\alpha_2^*\alpha_4^*\alpha_5^*} = a_{\alpha_1\alpha_3\alpha_2\alpha_1^*\alpha_3^*\alpha_2^*\alpha_4^*\alpha_5^*}$$
$$= a_{\beta_1\beta_2\beta_2\gamma_1^*} = a_{\beta_1\beta_2\beta_2\alpha_1^*\alpha_2^*} = a_{\alpha_1\alpha_1^*\beta_2\beta_1\alpha_2^*\alpha_3^*}$$

A starred lower index may be written as an unstarred upper index and an unstarred lower index as a starred upper index (but see 2.7). To avoid ambiguity I shall make the convention that all the upper indices are to be considered as occurring first and the lower ones afterwards. (Otherwise it would be necessary to stagger the indices, e.g. if $c = c_{\alpha_2}{}^{\alpha_1^*}$, then $c^{\alpha_1^*}{}_{\alpha_2}$ would be the transpose of a.)

As an example, let $\beta = \alpha\alpha^*$, $\gamma = \alpha^2$ and let $a_{\beta^3\gamma^*} = b^{\beta^3\gamma}_{\beta^3}$. Then we may write

$$b = b^{\beta_1\gamma_1}_{\beta_1\beta_2} = b^{\alpha_1\alpha_1^*\alpha_2\alpha_3}_{\beta_1\beta_2} = b^{\alpha_1\alpha_1^*\alpha_2\alpha_3}_{\alpha_2\alpha_4^*\alpha_3\alpha_5^*} = b^{\alpha_1\alpha_2\alpha_3\alpha_1^*}_{\alpha_2\alpha_3\alpha_4^*\alpha_5^*}$$

$$= a_{\beta_1^*\gamma_1^*\beta_1\beta_2} = a_{\alpha_1^*\alpha_1\alpha_2^*\alpha_3^*\beta_1\beta_2} = a_{\alpha_1^*\alpha_1\alpha_2^*\alpha_3^*\alpha_2\alpha_4^*\alpha_3\alpha_5^*} = a$$

Dummy composite indices may also be used to denote contraction between two sets of indices. If a tensor symbol has an upper $\overline{\beta}_i$ and a lower $\overline{\beta}_i$, this indicates contraction between the corresponding constituent upper and lower covariant and contravariant indices. Also a pair of lower indices $\overline{\beta}_i$ and $\overline{\beta}^*_i$ indicate a corresponding set of contractions between lower indices and similarly for a pair of upper indices. Thus, if a and b are as above,

$$b^{\overline{\beta}_1\gamma_1}_{\overline{\beta}_1\beta_1} = b^{\overline{\alpha}_1\overline{\alpha}_2^*\gamma_1}_{\overline{\alpha}_1\overline{\alpha}_2^*\alpha_1\alpha_1^*} = b^{\overline{\alpha}_1\overline{\alpha}_2^*\alpha_1\alpha_2}_{\overline{\alpha}_1\overline{\alpha}_2^*\alpha_1\alpha_3^*}$$

$$= a_{\overline{\beta}_1^*\gamma_1^*\overline{\beta}_1\beta_1} = a_{\overline{\alpha}_1^*\overline{\alpha}_2\alpha_1^*\alpha_2^*\overline{\alpha}_1\overline{\alpha}_2^*\alpha_1\alpha_3^*} = a_{\overline{\alpha}_2\overline{\alpha}_1\alpha_1\overline{\alpha}_1^*\alpha_1^*\alpha_2^*\overline{\alpha}_2^*\alpha_3^*}$$

If the numbering of an index type is not carried beyond 1, then no confusion will arise if this numerical suffix is omitted. For example, if a and b are as above, we may write

$$b = b^{\beta\gamma}_{\beta_1\beta_2} = b^{\alpha_1\alpha^*\alpha_2\alpha_3}_{\beta_1\beta_2} = a_{\beta^*\gamma^*\beta_1\beta_2} = a_{\alpha_1\alpha_2\alpha_3\alpha^*\beta_1\beta_2}$$

and

$$b^{\overline{\beta}_1\gamma_1}_{\overline{\beta}_1\beta_1} = b^{\overline{\beta}\gamma}_{\overline{\beta}\beta} = b^{\overline{\alpha}_1\overline{\alpha}_2^*\gamma}_{\overline{\alpha}_1\overline{\alpha}_2^*\alpha\alpha^*} = a_{\overline{\beta}^*\gamma^*\overline{\beta}\beta}$$

This notation agrees with that of 2.6.

It should be noted that while $b^{\beta\gamma}_{\beta_2\beta_1}$ denotes a transpose of b, the tensor $b^{\gamma\beta}_{\beta_1\beta_2}$ is just b again. Transposes are only obtained if differently

numbered indices of the same type are permuted. (This notation is dealt with more fully and with greater generality in (8), the first section of which treats tensors and the tensor operations from an entirely abstract algebraic point of view.)

2.9 Notation. In view of the convention made in 2.8 about upper indices having to be considered first, it is necessary to make a slight alteration in the definition of the product notation ab made in 2.3, when a and b possess covariant upper indices or contravariant lower indices. If $a = a_{\gamma}^{\beta}$ and $b = b_{\eta}^{\zeta}$, then the product $c_{\gamma\eta}^{\beta\zeta} = a_{\gamma}^{\beta} b_{\eta}^{\zeta}$ will be defined by

$$c^{\alpha_1 \cdots \alpha_{\lambda+u}\, \alpha_1^* \cdots \alpha_{\lambda'+u'}^*}_{\alpha_{\lambda'+u'+1} \cdots \alpha_{\lambda'+u'+v}^*\, \alpha_{\lambda+u+1}^* \cdots \alpha_{\lambda+u+d'+v'}^*} = a^{\alpha_1 \cdots \alpha_\lambda \alpha_1^* \cdots \alpha_{\lambda'}^*}_{\alpha_{\lambda'+u'+1} \cdots \alpha_{\lambda'+u'+d}^*\, \alpha_{\lambda+u+1}^* \cdots \alpha_{\lambda+u+d'}^*} b^{\alpha_{\lambda+1} \cdots \alpha_{\lambda+u}\, \alpha_{\lambda'+1}^* \cdots \alpha_{\lambda'+u'}^*}_{\alpha_{\lambda'+u'+d+1} \cdots \alpha_{\lambda'+u'+v}^*\, \alpha_{\lambda+u+d'+1}^* \cdots \alpha_{\lambda+u+d'+v'}^*}$$

where $\beta = \alpha^{\lambda} \alpha^{*\lambda'}$, $\gamma = \alpha^{d} \alpha^{*d'}$, $\zeta = \alpha^{u} \alpha^{*u'}$, $\eta = \alpha^{v} \alpha^{*v'}$. This is a transpose of the product given in 2.3, but I shall still use the notation c = ab. Since the symbols a and b are each associated with the particular division into upper and lower indices given, this does not lead to an ambiguity. With this notation all the upper indices are considered as occurring first and the lower ones afterwards. (The definition given in 2.3 would give rise to a symbol of the form $d^{\beta}{}_{\gamma}{}^{\zeta}{}_{\eta}$ with staggered indices. For many purposes, this symbol d would behave in the same way as c, but see 2.10.) The associative law of multiplication still holds with the new definition.

In view of the remark made at the end of 2.8, we observe that $c_{\eta\gamma}^{\zeta\beta} = c_{\gamma\eta}^{\beta\zeta}$, hence, with the notation employed here,

$$a_{\gamma}^{\beta} b_{\eta}^{\zeta} \neq b_{\eta}^{\zeta} a_{\gamma}^{\beta}$$

in general, since outer multiplication is not commutative. The second expression is a transpose of the first. However, it is correct to write

$$f^{\beta_1 \cdots \beta_\lambda}_{\gamma_1 \cdots \gamma_\lambda}\, g^{\beta_{\lambda+1} \cdots \beta_{\lambda+\mu}}_{\gamma_{\lambda+1} \cdots \gamma_{\lambda+\nu}} = g^{\beta_{\lambda+1} \cdots \beta_{\lambda+\mu}}_{\gamma_{\lambda+1} \cdots \gamma_{\lambda+\nu}}\, f^{\beta_1 \cdots \beta_\lambda}_{\gamma_1 \cdots \gamma_\lambda}$$

and the same would hold if the γ's were all replaced by β'_λ. The left hand side gives the correct transpose that fg is of gf.

2.10 Contracted products. Let $a = a^{\beta}_{\gamma}$ and $b = b^{\gamma}_{\delta}$. Then the contracted product c of a with b is given by

$$a \bullet b = a^{\beta}_{\bar{\gamma}}\, b^{\bar{\gamma}}_{\delta} = c^{\beta}_{\delta}$$

It will be seen that this definition depends on the particular division of the indices of a and b into upper and lower indices, as was the case to a lesser extent with the definition of outer product given in 2.9.

Assuming conformability in each case, we have

$$a \bullet (b \bullet c) = (a \bullet b) \bullet c \ ;$$

$$(a + b) \bullet c = a \bullet c + b \bullet c, \quad d \bullet (a + b) = d \bullet a + d \bullet b \ ;$$

$$(a \bullet b)(c \bullet d) = (ac) \bullet (bd).$$

The last of these relations depends on the definition given in 2.9 of the notation for outer product.

2.11 Notation. If a is any tensor and r is a non-negative integer, I shall use the notation

$$(a)^r$$

to denote the outer product aa.. a of r a's in succession, the bracket being used to avoid possible confusion of r with an upper tensor index.

If $r = 0$ the usual convention about zero exponents will be used, that is, $(a)^0$ denotes the scalar 1.

3. Some tensor systems.

Let $\mathfrak{R}$ be an algebraically closed field without characteristic which is of infinite transcendence over the rationals. (For many purposes, the only essential property of $\mathfrak{R}$ is that it is without characteristic.) Let $\mathfrak{C}$ denote the set of all tensors which may be represented by a set of components taken from $\mathfrak{R}$, the indices of the tensor components taking the values $1,..,n$. Thus each element of $\mathfrak{C}$ may be represented by (or, if desired, considered as) an ordered set of elements $a^{i_1 \cdots i_\alpha}_{j_1 \cdots j_\beta}$ of $\mathfrak{R}$, where $i_1,..,j_\alpha$ take the values $1,..,n$. The set $\mathfrak{C}$ is then an algebraic system closed with respect to the tensor operations 2.2, 2.3, 2.4 and 2.5.

3.1 Tensor subsystems of $\mathfrak{C}$. A subset $\mathfrak{T}$ of $\mathfrak{C}$ may also be closed with respect to the tensor operations. If, in addition, all the zeros of $\mathfrak{C}$ belong to $\mathfrak{T}$ and the scalars of $\mathfrak{C}$ which belong to $\mathfrak{T}$ constitute a subfield of $\mathfrak{R}$, then I shall call $\mathfrak{T}$ a tensor subsystem of $\mathfrak{C}$ or simply a tensor system. The scalars of $\mathfrak{T}$ may or may not be the whole field $\mathfrak{R}$. Thus $\mathfrak{C}$ is an example of a tensor system.

I shall use the notation $\mathfrak{T}_\beta$ where $\beta \in \mathfrak{G}$ for the set of elements of $\mathfrak{T}$ of type β (see 2.6). The set of scalars of $\mathfrak{T}$ is then the set $\mathfrak{T}_1$.

3.2 Units. The symbol I_α^α or $I_{(\alpha)\alpha}^{\ \ \alpha}$ will be used to denote that element of $\mathcal{C}$ whose components are given by the Kronecker delta δ_j^i. The symbol $I_{(\alpha^*)\alpha^*}^{\ \ \ \alpha^*}$ will denote the same tensor but with the roles of the upper and lower indices interchanged, i.e.

$$I_{(\alpha^*)} = I_{(\alpha^*)\alpha^*}^{\ \ \ \alpha^*} \doteq I_{(\alpha)\alpha}^{\ \ \alpha} = I_{(\alpha)}$$

More generally, let $\beta \in \mathcal{B}$, then $I_{(\beta)}$ denotes that element of $\mathcal{C}_{\beta\beta^*}$ given by

$$I_{(\beta)\beta}^{\ \ \beta} = (I_{(\alpha)})^{\wedge} (I_{(\alpha^*)})^{\Delta}$$

where $\beta = \alpha^{\wedge} \alpha^{*\Delta}$

From the properties of the Kronecker delta it follows that for each $a_\gamma^\beta \in \mathcal{C}$,

$$I_{(\beta)} \bullet a = a = a \bullet I_{(\gamma)}$$

The tensors $I_{(\beta)}$ are thus unit elements with respect to contracted products.

The tensor $I_{(\beta)}$ will be called the β-unit of $\mathcal{C}$. The $\beta^{\wedge}$-unit will be denoted by $I_{(\beta)}^{(\wedge)}$. We have

$$I_{(\beta)}^{(\wedge)} = I_{(\beta^{\wedge})} = (I_{(\beta)})^{\wedge}$$

and, more generally,

$$I_{(\beta\gamma)} = I_{(\beta)} I_{(\gamma)} = I_{(\gamma)} I_{(\beta)}$$

These properties follow from the fact that $I_{(\alpha)} I_{(\alpha^*)} = I_{(\alpha^*)} I_{(\alpha)}$ (which is true by virtue of the definition of the notation for outer product given in 2.9). I shall usually write $I^{(\wedge)}$ for $I_{(\alpha)}^{(\wedge)}$.

3.3 Rational tensors and invariant tensors. The system $\mathcal{R}$ of rational tensors will be defined to be the tensor system generated by the α-unit I.

Thus $\mathcal{R}$ is the smallest tensor subsystem of $\mathcal{C}$ which contains I. The elements of each set $\mathcal{R}_{\alpha\hat{}\alpha^*\hat{}}$ are the tensors of the form

$$A = \sum_k \lambda_k I^{\alpha_1}_{\alpha_{k(1)}} I^{\alpha_2}_{\alpha_{k(2)}} \ldots I^{\alpha_s}_{\alpha_{k(s)}} \qquad 3.3a$$

summed over all permutations k of $1,\ldots,r$, the λ'_k being rational numbers. If $r = 0$, A is simply a rational number, i.e. $\mathcal{R}_1$ is the rational subfield of $\mathfrak{R}$. Each set $\mathcal{R}_\beta$, for which β is not of the form $\alpha\hat{}\alpha^*\hat{}$, contains only the zero of $\mathfrak{R}$. It is clear from the definition of a tensor system, given in 3.1, that $\mathcal{R}$ contains all these elements. ($\mathcal{R}$ must contain all the rational elements of $\mathfrak{R}$ since they constitute the smallest subfield of $\mathfrak{R}$.) Conversely, this set of elements is closed with respect to the operations of addition, transposition, outer multiplication and contraction, as is easily seen. ($I^{\alpha}_{\alpha} = n$, an integer.) Since every unit $I_{(\beta)}$ of $\mathcal{C}$ is an outer product of tensors equal to I, the system $\mathcal{R}$ contains all the units of $\mathcal{C}$.

The system $\mathcal{I}$ of invariant tensors is the system generated by I and the elements of $\mathfrak{R}$. The elements of each set $\mathcal{I}_{\alpha\hat{}\alpha^*\hat{}}$ are the tensors of the form 3.3a, but where the λ'_k can be any elements of $\mathfrak{R}$. The field $\mathcal{I}_1$ is just the field $\mathfrak{R}$. Each set $\mathcal{I}_\beta$, where β is not of the form $\alpha\hat{}\alpha^*\hat{}$, contains only the zero of $\mathcal{C}_\beta$. These facts follow as do those for $\mathcal{R}$ above. $\mathcal{R}$ is a subsystem of $\mathcal{I}$ and so $\mathcal{I}$ also contains all the units of $\mathcal{C}$. It will be seen that the components of any element of $\mathcal{I}$ are independent of the coordinate system, since this applies in particular to the tensor I. (For the converse see 8.8.)

3.4 Substitutional operators. There is a close relationship between the elements of $\mathcal{J}$ and substitutional expressions (elements of the group ring of a symmetric group). The element A of $\mathcal{J}$ defined by 3.3a may be associated with the substitutional expression $\sum_k \lambda_k k$ (the k's are permutations and the λ_k may be any elements of $\mathfrak{K}$). The operation of multiplication between substitutional expressions corresponds to a contracted product between the elements of $\mathcal{J}$, as is easily verified. However, this correspondence is not exact because of the existence of certain identical relations satisfied by I (see 3.7a). Furthermore, the system $\mathcal{J}$ also admits the operations of contraction and outer multiplication.

It is possible to associate a substitutional expression with an element of $\mathcal{J}$ in other ways also. For example, if $\beta \in \mathfrak{S}$ and

$$A_{(\beta)} = \sum_k \lambda_k I_{(\beta)\beta_{k(1)}}^{\beta_1} I_{(\beta)\beta_{k(2)}}^{\beta_2} \dots I_{(\beta)\beta_\nu}^{\beta_\nu}$$

then $A_\beta \in \mathcal{J}$. Addition and multiplication between substitutional expressions again correspond to addition and contracted multiplication between the associated elements of $\mathcal{J}$. For a particular β, however, only some of the elements of the sets $\mathcal{J}_{\beta^\nu \beta^{*\nu}}$ will correspond to substitutional expressions in this way.

3.5 Symmetrizers and skew-symmetrizers. Let $\beta \in \mathfrak{S}$ and define

$$P_{(\beta)}^{(\nu)} = \frac{1}{\nu!} \sum_k I_{(\beta)\beta_{k(1)}}^{\beta_1} I_{(\beta)\beta_{k(2)}}^{\beta_2} \dots I_{(\beta)\beta_{k(\nu)}}^{\beta_\nu},$$

$$N_{(\beta)}^{(\nu)} = \frac{1}{\nu!} \sum_k \eta_k I_{(\beta)\beta_{k(1)}}^{\beta_1} I_{(\beta)\beta_{k(2)}}^{\beta_2} \dots I_{(\beta)\beta_{k(\nu)}}^{\beta_\nu}$$

where $\eta_k = +1$ or -1 according as k is an even or odd permutation of

CHAPTER 7

1,..,r. The tensors $P^{(\lambda)}_{(\beta)}$ and $N^{(\lambda)}_{(\beta)}$ are elements of $\mathfrak{R}$ and correspond, respectively, to the substitutional operators "the positive symmetric group on r symbols" and "the negative symmetric group on r symbols" (see (9) p. 8). I shall usually write $P^{(\lambda)}$ to denote $P^{(\lambda)}_{(\kappa)}$ and $N^{(\lambda)}$ to denote $N^{(\lambda)}_{(\alpha)}$.

The tensors $P^{(\lambda)}_{(\beta)}$ and $N^{(\lambda)}_{(\beta)}$ are important because they may be used to express symmetry or skew-symmetry of a tensor. Let $a^{\gamma}_{\beta_1 \cdots \beta_\lambda}$, $b^{\beta_1 \cdots \beta_\lambda}_{\gamma} \in \mathfrak{C}$ Then necessary and sufficient conditions for a and b to be symmetric in their β-indices are respectively

$$a \bullet P^{(\lambda)}_{(\beta)} = a, \quad P^{(\lambda)}_{(\beta)} \bullet b = b.$$

The corresponding conditions for skew-symmetry are

$$a \bullet N^{(\lambda)}_{(\beta)} = a, \quad N^{(\lambda)}_{(\beta)} \bullet b = b.$$

(The tensor $a \bullet P^{(\lambda)}_{(\beta)}$ is the tensor a symmetrized with respect to its lower β-indices, etc.)

3.6 Dimension of an index type. I define the dimension of the index type β to be the integer

$$\nu(\beta) = I_{(\beta)\overline{\beta}}^{\overline{\beta}}$$

We have

$$\nu(\alpha\beta) = \nu(\alpha)\nu(\beta) \text{ and } \quad \nu(\beta^\lambda) = \{\nu(\beta)\}^\lambda$$

(see 3.2). Hence

$$\nu(\alpha^\lambda \alpha^{*\mu}) = n^{\lambda+\mu}$$

Each set $\mathfrak{C}_\beta$ is a vector space over $\mathfrak{R}$ and the dimension of this

vector space is $\nu(\beta)$. The vector space $\mathcal{C}_{\alpha^r\alpha^{*s}}$ is the tensor product of r vector spaces $\mathcal{C}_\alpha$ and s vector spaces $\mathcal{C}_{\alpha^*}$, each of which is n-dimensional.

3.7 Identities involving $N^{(\lambda)}_{(\beta)}$. The components of the tensor $r!\,N^{(\lambda)}_{(\beta)}$ are given by the generalized Kronecker delta

$$\delta^{i_1\cdots i_\lambda}_{j_1\cdots j_\lambda} = \begin{vmatrix} \delta^{i_1}_{j_1} & \delta^{i_1}_{j_2} & \cdots & \delta^{i_1}_{j_\lambda} \\ \delta^{i_2}_{j_1} & & & \vdots \\ \vdots & & & \\ \delta^{i_\lambda}_{j_1} & \cdots & & \delta^{i_\lambda}_{j_\lambda} \end{vmatrix}$$

(see (6) p. 32). If $r > n$, $\delta^{i_1\cdots i_\lambda}_{j_1\cdots j_\lambda} = 0$ for all $i_1,\ldots,j_\lambda$, since the i's and j's cannot all be distinct. Thus $N^{(\lambda)} = 0$ for all $r > n$ and in particular $N^{(\lambda+1)} = 0$. Similarly, for each $\beta \in \mathfrak{S}$, we have

$$N^{(\nu(\beta)+1)}_{(\beta)} = 0 \qquad 3.7a$$

since the β-indices of $N^{(\lambda)}_{(\beta)}$ are $\nu(\beta)$-dimensional.

Now the following relation follows from a substitutional identity (see (9) p. 10)

$$(r+1)\; N^{(\lambda+1)}_{(\beta)} = N^{(\lambda)}_{(\beta)} I_{(\beta)} - \sum_{i=1}^{\lambda} N^{(\lambda)\,\beta_1\beta_2\cdots\beta_\lambda}_{(\beta)\,\beta_1\cdots\beta_{i-1}\beta_{\lambda+1}\beta_{i+1}\cdots\beta_\lambda} I^{\beta_{\lambda+1}}_{(\beta)\,\beta_i} \qquad 3.7b$$

Hence $(r+1)\; N^{(\lambda+1)\beta_1\cdots\beta_\lambda\overline{\beta}}_{(\beta)\,\beta_1\cdots\beta_\lambda\overline{\beta}} = N^{(\lambda)}_{(\beta)}\nu(\beta) - r\,N^{(\lambda)}_{(\beta)}$, i.e.

$$N^{(\lambda+1)\beta_1\cdots\beta_\lambda\overline{\beta}}_{(\beta)\,\beta_1\cdots\beta_\lambda\overline{\beta}} = \frac{\nu(\beta)-\lambda}{\lambda+1} N^{(\lambda)}_{(\beta)} \qquad 3.7c$$

By repeatedly applying 3.7c with $\beta = \alpha$, $\nu(\beta) = n$ and $r = n-1, n-2, \ldots, 1$, we get

$$N^{(n)\,\alpha\,\overline{\alpha}_{n-1}\overline{\alpha}_{n-2}\cdots\overline{\alpha}_1}_{\alpha\,\overline{\alpha}_{n-1}\overline{\alpha}_{n-2}\cdots\overline{\alpha}_1} = \frac{1}{n} I \qquad 3.7d$$

since $\frac{1}{n}\cdot\frac{2}{n-1}\cdot\ldots\cdot\frac{n-1}{2} = \frac{1}{n}$ and $N^{(1)} = I$. From 3.7b and 3.7a, if $\nu = \nu(\beta)$ we get

$$N^{(\nu)}_{(\beta)} I_{(\beta)} = \sum_{i=1}^{\nu} N^{(\nu)\,\beta_1\beta_2\cdots\beta_\nu}_{(\beta)\,\beta_1\cdots\beta_{i-1}\beta_{\nu+1}\beta_{i+1}\cdots\beta_\nu} I^{\beta_{\nu+1}}_{(\beta)\,\beta_i} \qquad 3.7e$$

CHAPTER 7

These relations will be of importance later.

3.8 Alternating tensors. The tensors $\epsilon = \epsilon^{\alpha_1 \cdots \alpha_m}$ and $\epsilon^* = \epsilon^*_{\alpha_1 \cdots \alpha_m}$ will be defined as tensors whose components in some fixed frame of reference are given by

$$\epsilon^{i_1 \cdots i_m} = \epsilon^*_{i_1 \cdots i_m} = \begin{cases} +1 \text{ if } i_1, \ldots i_m \text{ is an even permutation of } 1, \ldots, n \\ -1 \text{ if } i_1, \ldots i_m \text{ is an odd permutation of } 1, \ldots, n \\ 0 \text{ otherwise} \end{cases}$$

The components of ϵ and ϵ^* are relatively invariant under change of coordinate system (see (6) p 320). Thus, the definitions of ϵ and of ϵ^* are determined up to proportionality independently of the frame of reference.

Since $\epsilon^{i_1 \cdots i_m} \epsilon^*_{j_1 \cdots j_m} = \delta^{i_1 \cdots i_m}_{j_1 \cdots j_m}$ (see (6) p. 320), we have

$$\epsilon \epsilon^* = n! N^{(m)} \qquad 3.8a$$

Hence, by 3.7d, (or (6) p. 320)

$$\epsilon^*_{\alpha \bar{\alpha}_1 \cdots \bar{\alpha}_{m-1}} \epsilon^{\alpha \bar{\alpha}_1 \cdots \bar{\alpha}_{m-1}} = (n-1)! I \qquad 3.8b$$

3.9 The systems $\mathfrak{R}^+$ and $\mathfrak{F}^+$. The system $\mathfrak{R}^+$ is the tensor system generated by ϵ, ϵ^* and the elements of $\mathfrak{R}$. Similarly, $\mathfrak{F}^+$ is the tensor system generated by ϵ, ϵ^* and the elements of $\mathfrak{F}$. The non-zero elements of $\mathfrak{R}^+$ and $\mathfrak{F}^+$ belong to the sets $\mathfrak{R}^+_\beta$ and $\mathfrak{F}^+_\beta$, respectively, for which $\beta = \alpha \wedge \alpha^{*\wedge r+m}$ or $\beta = \alpha^{*\wedge m+s} \alpha^{*\wedge}$ $(r, s \geq 0)$ since ϵ is of type $\alpha^{*\wedge m}$ and ϵ^* is of type $\alpha^{\wedge m}$. Each element of $\mathfrak{R}^+$ or of $\mathfrak{F}^+$ can be written as a sum of terms each of which is homogeneous in ϵ and in ϵ^*. But, by 3.8a, the product of ϵ with ϵ^* is an element of $\mathfrak{R}$. Thus, each term of the sum involves

only $(\epsilon)^s$ or $(\epsilon^*)^s$ for some s. Since the tensors of a sum must all have the same type each term of the sum must involve ϵ to the same power s or ϵ^* to the same power s. It follows that every element of $\mathcal{R}^+_\beta$ can be written in the form

$$A \cdot (\epsilon)^s \qquad \text{or} \qquad (\epsilon^*)^s \cdot B \tag{3.9a}$$

where $A^{\beta^*}_{\alpha^s}$, $B^{\alpha^s}_{\beta} \in \mathcal{R}$ and the corresponding result holds for $\mathcal{J}^+$. The first of these expressions will be said to be an element of <u>weight</u> s and the second, an element of weight -s. The weight of the product of two elements is the sum of their weights. The elements of $\mathcal{J}^+$ of weight zero are the elements of $\mathcal{J}$ and similarly for $\mathcal{R}^+$, $\mathcal{R}$. Hence, for every element a of there exists an element b of $\mathcal{C}$ such that

$$a\ b\ \epsilon\ \mathcal{R}$$

and the corresponding result holds for $\mathcal{J}^+$. The significance of this fact will be seen in 6.7. This is also a defining property of $\mathcal{R}^+$ and $\mathcal{J}^+$ as will be proved in 7.8.

3.10 <u>Young idempotents</u>. If $\lambda = (\lambda_1, .., \lambda_p)$ is a partition of the integer $m > 0$ and $\mu = (\mu_1, .., \mu_q)$ is the conjugate partition to λ then irreducible idempotent substitutional operators corresponding to λ may be constructed by means of Young tableaux (see (9) p. 11.). These will correspond to certain elements of $\mathcal{R}$. Thus, such an idempotent $E^{\{\lambda\}\alpha^m}_{\alpha^m}$ may be constructed by putting $E^{\{\lambda\}} = \kappa^{\{\lambda\}} P^{\{\lambda\}} \cdot N^{\{\lambda\}}$ where $P^{\{\lambda\}} = P^{(\lambda_1)} P^{(\lambda_2)} .. P^{(\lambda_p)}$ and where $N^{\{\lambda\}}$ is a suitable transpose of $N^{(\mu_1)} N^{(\mu_2)} .. N^{(\mu_q)}$ $\kappa^{\{\lambda\}}$ being a certain rational number depending on λ. We have $E^{\{\lambda\}} \cdot E^{\{\lambda\}} = E^{\{\lambda\}}$.

If λ is a partition into more than n parts, then $\mu_i > n$ for some i so that, by 3.7a, $E^{\{\lambda\}} = 0$. Also, since $N^{(n)} = n!\epsilon\epsilon^*$ we have

$$E^{\{\lambda^n\}} = a\ b \qquad 3.10a$$

for some $a^{\alpha m}$, $b_{\alpha m} \in \mathcal{R}^+_n$. (b is simply a transpose of $(\epsilon^*)^\sim$.)

4. Forms and invariant theory.

The tensors considered in section 3 had sets of components which were essentially independent of the choice of coordinate system. For a general element of $\mathcal{C}$, however, this is clearly not the case. The set of components of a tensor is a function not only of the tensor itself but also of the coordinate system used. Since I am concerned here only with the formal algebraic properties of tensors with respect to the tensor operations, it is desirable that the set of components of a tensor should be obtainable from that tensor and the frame of reference by means of tensor operations. It will be necessary to introduce components of a tensor, so as to be able to relate tensor algebra to the ordinary algebra of geometry.

4.1 Frames. An ordered set $x^\alpha_{(1)}, \ldots, x^\alpha_{(n)}$; $u_{(1)\alpha}, \ldots, u_{(n)\alpha}$ of elements of will be said to constitute a frame $\mathcal{F}$ for $\mathcal{C}$ if

$$u_{(i)} \cdot x_{(j)} = \delta_{ij} \qquad 4.1a$$

and

$$\sum_{i=1}^{n} x_{(i)}\ u_{(i)} = I \qquad 4.1b$$

It follows from 4.1a and 4.1b by the usual arguments concerning bases for vector spaces that the sets $x_{(1)},..,x_{(n)}$ and $u_{(1)},..,u_{(n)}$ are linearly independent sets spanning $\mathcal{C}_{\alpha^*}$ and $\mathcal{C}_{\alpha}$, respectively. Furthermore, $x_{(1)},..,x_{(m)}$ can be any set of n linearly independent elements of $\mathcal{C}_{\alpha^*}$. The corresponding u's are then determined and are given by

$$u_{(i)} = \frac{1}{\epsilon^*\bullet(x_{(1)}..x_{(m)})}\,\epsilon^*\bullet(x_{(1)}..x_{(i-1)}\,I\,x_{(i+1)}..x_{(m)}) \qquad 4.1c$$

4.2 Components of a tensor in a given frame. Let $a^{\alpha_1..\alpha_n}_{\alpha_1..\alpha_s} \in \mathcal{C}$, let $\mathcal{F}$ be as in 4.1 and define

$$a^{i_1..i_n}_{j_1..j_s} = (u_{(i_1)}..u_{(i_n)})\bullet a\bullet(x_{(j_1)}..x_{(j_s)}) \qquad 4.2a$$

Then the set of scalars $a^{i_1..i_n}_{j_1..j_s}$ are the components of a in the frame $\mathcal{F}$. It is also possible to express the tensor a in terms of the elements of $\mathcal{F}$ and the components of a in $\mathcal{F}$, for by 4.2a and 4.1b,

$$a = \sum_{\substack{i_1..i_n=1\\ j_1..j_s=1}}^{m} (x_{(i_1)}..x_{(i_n)})\; a^{i_1..i_n}_{j_1..j_s}\; (u_{(j_1)}..u_{(j_s)}). \qquad 4.2b$$

The components of the element $x_{(i)}$ of $\mathcal{F}$ are 0,.., 0, 1, 0, .., 0 with 1 in the i^{th} place, by 4.1b. The same holds for $u_{(i)}$.

4.3 Forms. Let $x^1,..,x^n$ be a set of n contravariant variables. Let $\phi \equiv \phi(x^1,..,x^n)$ be a form, homogeneous of degree r in $x^1,..,x^n$ then we can write

$$\phi \equiv \sum_{i_1..i_n=1}^{n} a_{i_1..i_n}\, x^{i_1}..x^{i_n}$$

Thus, if x^i and $a_{i_1..i_n}$ are the sets of components in a frame $\mathcal{F}$ of the tensors x^α and $a_{\alpha_1..\alpha_n}$ respectively, we have

CHAPTER 7

$$\phi \equiv \sum_{i_1,..,i_\lambda} a\bullet(x_{(i_1)}..x_{(i_\lambda)})(u_{(i_1)}..u_{(i_\lambda)})\bullet(x)^\lambda \equiv a\bullet(x)^\lambda$$

by 4.2a and 4.1b ($x_{(1)},..,x_{(n)}$; $u_{(1)},..,u_{(m)}$ are as in 4.1). Similarly if ψ is a form of degree s in the n covariant variables $u_1,..u_m$ which are the components in $\mathcal{F}$ of the tensor u_α, we can write $\psi \equiv (u)^\delta\bullet b$ for some $b^{\alpha^\delta} \in \mathcal{C}$. More generally if χ is a form, respectively homogeneous of degree $r_1,..,r_m$ in the components in $\mathcal{F}$ of the tensors $y^\beta_{(1)},..,y^\delta_{(m)}$, then we can write

$$\chi = c\bullet\{(y_{(1)})^{\lambda_1}..(y_{(m)})^{\lambda_m}\}.$$

The form $\phi \equiv a\bullet(x)^\lambda$ is identically zero if and only if the symmetric part of a vanishes, i.e. if and only if $a\bullet P^{(\lambda)} = 0$ (see 3.5). Similarly the condition for $\psi \equiv (u)^\alpha\bullet b$ to be identically zero is $P^{(\alpha)}\bullet b = 0$. The forms ϕ and ψ correspond uniquely to the respective symmetric parts of a and b, i.e. to $a\bullet P^{(\lambda)}$ and $P^{(\alpha)}\bullet b$. More generally, the form χ given above uniquely determines and is determined by the tensor

$$c\bullet\{P^{(\lambda_1)}_{(\beta)}..P^{(\lambda_m)}_{(\delta)}\}$$

and this tensor vanishes if and only if $\chi \equiv 0$. A corresponding result would also hold if the products were written in the reverse order.

4.4 Clebsch forms. Let $z^\alpha_{(1)},..,z^\alpha_{(m)}$ be n variable elements of $\mathcal{C}_{\alpha^*}$ and put $y^{(\lambda)} = N^{(\lambda)}\bullet(z_{(1)}..z_{(\lambda)})$, $(r = 1,..,n)$. Then the components of $y^{(\lambda)}$ in a frame $\mathcal{F}$ are given by

$$y^{(\lambda)i_1..i_\lambda} = \frac{1}{\lambda!}\begin{vmatrix} z^{i_1}_{(1)} & .. & z^{i}_{(\lambda)} \\ \vdots & & \vdots \\ z^{i_\lambda}_{(1)} & .. & z^{i_\lambda}_{(\lambda)} \end{vmatrix}$$

where $z^i_{(a)}$ are the components in $\mathfrak{F}$ of $z_{(a)}$ (see 3.7). Let $\lambda = (\lambda_1, .., \lambda_p)$ be a partition of the integer $m > 0$ into not more than n parts and let $\mu = (\mu_1, .., \mu_q)$ be the partition conjugate to λ. Then the tensor

$$v^{\{\lambda\}} = y^{(\mu_1)} y^{(\mu_2)} .. y^{(\mu_q)}$$

is called a Clebsch variable tensor of type $\{\lambda\}$ (see (6) p.320). A form $b \cdot v^{\{\lambda\}}$, where $b_{\alpha m} \in \mathcal{C}$, is called a Clebsch form of type $\{\lambda\}$. We have

$$\begin{aligned} b \cdot v^{\{\lambda\}} &= b \cdot \{y^{(\mu_1)} .. y^{(\mu_q)}\} = b \cdot \{(N^{(\mu_1)} \cdot z_{(1)} .. z_{(\mu_1)}) .. (N^{(\mu_q)} \cdot z_{(1)} .. z_{(\mu_q)})\} \\ &= b \cdot (N^{(\mu_1)} N^{(\mu_2)} .. N^{(\mu_q)}) \cdot J \cdot \{(z_{(1)})^{\lambda_1} .. (z_{(p)})^{\lambda_p}\} \\ &= b \cdot J \cdot N^{\{\lambda\}} \cdot \{(z_{(1)})^{\lambda_1} .. (z_{(p)})^{\lambda_p}\} \end{aligned} \qquad 4.4a$$

where J is a suitable permutation (transpose of $I^{(m)}$) and where $N^{\{\lambda\}}$ is as in 3.10.

If $\phi(z_{(1)}, .., z_{(p)}) \equiv a \cdot \{(z_{(1)})^{\lambda_1} .. (z_{(p)})^{\lambda_p}\}$ is identically equal to a Clebsch form of type $\{\lambda\}$ (in $z_{(1)}, \ldots, z_{(p)}$) then it follows that

$$\phi(z_{(1)}, .., z_{(p)}) \equiv a \cdot E^{\{\lambda\}} \cdot \{(z_{(1)})^{\lambda_1} .. (z_{(p)})^{\lambda_p}\} \qquad 4.4b$$

where $E^{\{\lambda\}}$ is as in 3.10. For if $\phi \equiv b \cdot v^{\{\lambda\}}$, then by 4.4a and 4.3, we have

$$a \cdot \{P^{(\lambda_1)} .. P^{(\lambda_p)}\} = b \cdot J \cdot N^{\{\lambda\}} \cdot \{P^{(\lambda_1)} .. P^{(\lambda_p)}\}$$

i.e. $a \cdot P^{\{\lambda\}} = b \cdot J \cdot N^{\{\lambda\}} \cdot P^{\{\lambda\}}$ where $P^{\{\lambda\}}$ is as in 3.10. Now $K^{\{\lambda\}} N^{\{\lambda\}} \cdot P^{\{\lambda\}}$, like $E^{\{\lambda\}}$ $(= K^{\{\lambda\}} P^{\{\lambda\}} \cdot N^{\{\lambda\}})$, is an idempotent. Thus

$$K^{\{\lambda\}} b \cdot J \cdot N^{\{\lambda\}} \cdot P^{\{\lambda\}} = (K^{\{\lambda\}})^2 (b \cdot J \cdot N^{\{\lambda\}} \cdot P^{\{\lambda\}}) \cdot N^{\{\lambda\}} \cdot P^{\{\lambda\}}$$

whence

$$a \cdot P^{\{\lambda\}} = K^{\{\lambda\}} (a \cdot P^{\{\lambda\}}) \cdot N^{\{\lambda\}} \cdot P^{\{\lambda\}} = a \cdot E^{\{\lambda\}} \cdot P^{\{\lambda\}}$$

so that 4.4b follows.

The results of classical invariant theory are generally stated in terms of Clebsch forms (or their equivalents) since they are irreducible in a group theoretical sense. I shall require some of these basic results of invariant theory for the proofs of later results, but since the stated form of these basic results is not quite adequate for my purposes, independent proofs will be given based on the ideas developed here. I first prove two lemmas.

4.5 Lemma. Let a^{β}_{γ}, b^{β}_{δ}, $x_{\beta} \in \mathcal{C}$, where neither a nor b is zero. Then a necessary and sufficient condition for both the relations $x\bullet a = 0$, $x\bullet b = 0$ to hold is that $x\bullet c = 0$ where

$$c^{\beta_1}_{\beta_2^*\gamma\delta} \doteq P^{(2)\beta_1\beta_2}_{(\beta)\bar{\beta}_1\bar{\beta}_2}\, a^{\bar{\beta}_1}_{\gamma}\, b^{\bar{\beta}_2}_{\delta} = P^{(2)}_{(\beta)}\bullet(ab)$$

Proof. We have $x\bullet c = x_{\bar{\beta}}\, c^{\bar{\beta}}_{\beta^*\gamma\delta} = x_{\bar{\beta}_1}\, P^{(2)\bar{\alpha}_1\beta}_{(\alpha)\bar{\beta}_2\bar{\beta}_3}\, a^{\bar{\alpha}_2}_{\gamma}\, b^{\bar{\beta}_3}_{\delta}$ and this tensor is equal to $(x\, I_{(\beta)})\bullet P^{(2)}_{(\beta)}\bullet(ab)$. Now if $x\bullet a = 0$ and $x\bullet b = 0$, we have (see 3.5)

$$\begin{aligned} 2(x\, I_{(\beta)})\bullet P^{(2)}_{(\beta)}\bullet(ab) &= (x\, I_{(\beta)})\bullet(ab) + (I_{(\beta)}x)\bullet(ab) \\ &= (x\bullet a)b + a(x\bullet b) = 0 \end{aligned}$$

so that $x\bullet c = 0$. Conversely if $x\bullet c = 0$, then $(x\bullet a)b + a(x\bullet b) = 0$. Hence,

$$0 = x\bullet\{(x\bullet a)b + a(x\bullet b)\} = 2(x\bullet a)(x\bullet b)$$

That is, either $x\bullet a = 0$ or $x\bullet b = 0$. Suppose $x\bullet a = 0$ then

$$0 = (x\bullet a)b + a(x\bullet b) = a(x\bullet b)$$

so that $x\bullet b = 0$ since $a \neq 0$ (see 2.3). Similarly if $x\bullet b = 0$ then $x\bullet a = 0$. This proves the result.

4.6 **Lemma.** Let $\mathcal{Q}$ be a subclass of tensors of $\mathcal{C}_\beta$, let $\mathcal{F}$ be a frame for $\mathcal{C}$ and let $\phi_i(x,\mathcal{F}) = 0$ $(i = 1,.., m)$ be a set of linear homogeneous relations connecting the components in $\mathcal{F}$ of a tensor x_β. Suppose these relations represent a condition on elements of $\mathcal{Q}$ which is independent of the choice of $\mathcal{F}$. Then they are equivalent, for elements of $\mathcal{Q}$, to a relation

$$x \bullet F = 0$$

where $F^\beta_\gamma \in \mathcal{J}$ and where $F^\beta_\gamma \in \mathcal{R}$ if the $\phi_i(x)$ have rational coefficients. If the given set of relations consists of a single relation only $(m = 1)$, then there exists an element $f^\beta \in \mathcal{J}^+$, with $f^\beta \in \mathcal{R}^+$ if the relation has rational coefficients such that

$$\phi_1(x) = \xi\, x \bullet f$$

for all $x \in \mathcal{Q}$, where ξ is an element of $\mathcal{A}$ depending on $\mathcal{F}$.

Proof. In accordance with 4.3 the relations can be written as

$$x \bullet f_{(i)} = 0 \quad (i = 1,.., m)$$

where $f^\beta_{(i)} \in \mathcal{C}_{\beta^*}$, the components of each $f_{(i)}$ in $\mathcal{F}$ giving the coefficients in one of the relations. Let $\beta = \alpha^p \alpha^{*q}$, then

$$x_\beta = x_{\alpha_1 \ldots \alpha_p \alpha^*_1 \ldots \alpha^*_q} = x_{\alpha_1 \ldots \alpha_p \bar{\alpha}^*_1 \ldots \bar{\alpha}^*_q} I^{\alpha_1}_{\bar{\alpha}_1} \ldots I^{\alpha_q}_{\bar{\alpha}_q},$$

so that replacing each I by $\frac{1}{(n-1)!} \epsilon^*_{\alpha \bar{\alpha}_1 \ldots \bar{\alpha}_{n-1}} \epsilon^{\alpha \bar{\alpha}_1 \ldots \bar{\alpha}_{n-1}}$ (using 3.8b) we get

$$x = x \bullet a \bullet b \qquad 4.6b$$

$a^\beta_{\alpha^\wedge}$ being an element of $\mathcal{R}^+$ involving the ϵ^*'s $(r = p + q(n-1))$ and $b^{\alpha^\wedge}_\beta$ being an element of $\mathcal{R}^+$ involving the ϵ's. Thus $x \bullet a$ is an entirely covariant tensor which is obtained from x by "lowering" the contravariant

indices of x by means of ϵ^*. Putting $g_{(i)} = b \bullet f_{(i)}$ and using 4.6b, we can write 4.6a as

$$x \bullet a \bullet g_{(i)} = 0 \qquad (i = 1,..,m).$$

Let the components in $\mathfrak{F}$ of $g_{(i)}$ be $g_{(i)}^{i_1 \cdots i_\lambda}$ ($g_{(i)}$ is entirely contravariant). The $g_{(i)}^{i_1 \cdots i_\lambda}$ may be taken to be rational numbers if the original relations have rational coefficients since the components of ϵ and $f_{(i)}$, in $\mathfrak{F}$, can be taken as rational. We have

$$g_{(i)} = \sum_{i_1, \cdots, i_\lambda} g_{(i)}^{i_1 \cdots i_\lambda} \; x_{(i_1)} \cdots x_{(i_\lambda)}$$

where $x_{(1)}^{\alpha}, .., x_{(m)}^{\alpha} \in \mathfrak{F}$. Collecting terms together in groups homogeneous in $x_{(1)}, .., x_{(m)}$ we get

$$g_{(i)} = \sum_{j_1 + \cdots + j_m = \lambda} A_{(i; j_1, \cdots, j_m)} \bullet \{ (x_{(1)})^{j_1} \cdots (x_{(m)})^{j_m} \}$$

where $A_{(i;\ldots)}$ is some element of $\mathfrak{F}$. ($A_{(i;\ldots)}$ is a sum of permutations with the coefficients $g_{(i)}^{i_1 \cdots i_\lambda}$ attached.) If the relations $\phi_i(x) = 0$ had rational coefficients then $A_{(\ldots)} \in \mathfrak{R}$. Relations 4.6a now become

$$x \bullet a \bullet \Big[\sum_{j_1 + \cdots + j_m = \lambda} A_{(i; j_1, \cdots, j_m)} \bullet \{ (x_{(1)})^{j_1} \cdots (x_{(m)})^{j_m} \} \Big] = 0 \qquad (i = 1,.., m).$$

This set of equations is to represent the same condition on an element x of $\mathfrak{L}$, for each choice of $\mathfrak{F}$. This must hold in particular if $x_{(i)}$ is replaced by $\mu_i x_{(i)}$ with $\mu_i \neq 0$ so that we may assume that for each i, $A_{(i; j_1, \cdots, j_m)}$ is non-zero for only one set of $j_1, \cdots, j_m$ (i.e. the relations are isobaric). The relations may be written

$$x \bullet a \bullet A_{(i)} \bullet \{ (x_{(1)})^{j_{1i}} \cdots (x_{(m)})^{j_{mi}} \} = 0 \quad . \quad (i = 1,.., m). \qquad 4.6c$$

These must hold for each choice of $\mathfrak{F}$, that is, for each choice of

linearly independent $x_{(1)}^{\alpha}, \ldots, x_{(m)}^{\alpha}$. The relations may thus be replaced by the condition that 4.6c should hold identically in $x_{(1)}, \ldots, x_{(m)}$, i.e. that

$$x \bullet a \bullet A_{(i)} \bullet \{P^{(j_1 i)} \ldots P^{(j_m i)}\} = 0 \qquad (i = 1, \ldots, m)$$

(see 4.3). These relations can be combined in pairs by means of lemma 4.5 so as finally to obtain a single equivalent equation

$$x \bullet h = 0$$

where h is built up from the $A_{(i)}$'s, $P^{(j_1 j_2)}$'s, a and $P^{(2)}_{(\alpha\gamma)}$. Thus $h \in \mathcal{P}^+$ and if the relations $\phi_i(x) = 0$ have rational coefficients $h \in \mathcal{R}^+$. Now by 3.9, an element c_ξ of $\mathcal{C}$ may be chosen so that hc is an element of $\mathcal{P}$ or $\mathcal{R}$ according as $h \in \mathcal{P}^+$ or $\mathcal{R}^+$. Thus if $F = hc$, the equation is equivalent to $x \bullet F = 0$ and the first part of the lemma is proved.

If the given relations consist of the single equation $\phi_1(x) = 0$, this leads to a single relation 4.6c (m = 1). Now, for any fixed x of and for varying $x_{(1)}, \ldots, x_{(m)}$, the expression $x \bullet a \bullet A_{(1)} \bullet \{(x_{(1)})^{j_1} \ldots (x_{(m)})^{j_m}\}$ either vanishes identically in $x_{(1)}, \ldots, x_{(m)}$ or else it cannot vanish unless $x_{(1)}, \ldots, x_{(m)}$ become linearly dependent (only in this case do $x_{(1)}, \ldots, x_{(m)}$ not belong to some frame). But the condition that $x_{(1)}, \ldots, x_{(m)}$ should be linearly dependent is a single irreducible algebraic relation connecting the components of $x_{(1)}, \ldots, x_{(m)}$ (in some fixed frame) and is given by $\epsilon^*_\bullet(x_{(1)} \ldots x_{(m)}) = 0$ (since this is the condition that the determinant of these vectors should vanish). Hence

$$x \bullet a \bullet A_{(1)} \bullet \{(x_{(1)})^{j_1} \ldots (x_{(m)})^{j_m}\} \equiv \zeta \{\epsilon^*_\bullet(x_{(1)} \ldots x_{(m)})\}^{j}$$

holds identically in $x_{(1)},..,\ x_{(m)}$ for some j and some $\xi \in \mathfrak{R}$. Unless the relation $\phi_1(x) = 0$ is vacuous for elements x of $\mathfrak{Q}$ (in which case the lemma is trivially satisfied with $f^{\beta} = 0$), we must have $\xi \neq 0$ for some $x \in \mathfrak{Q}$, whence $j_1 = .. = j_m = j$. Thus

$$x\bullet a\bullet A_{(1)}\bullet \{(x_{(1)})^{\dot{j}} .. \ (x_{(m)})^{\dot{j}}\} \equiv \xi\{\epsilon^{*}_{\dot{j}}(x_{(1)}.. \ x_{(m)})^{\dot{j}}\}$$

Since $\epsilon^{*} = \epsilon^{*}\bullet N^{(m)}$, the left hand side of this is a Clebsch form of type $\{j^m\}$ in $x_{(1)},..,\ x_{(m)}$ (see 4.4). Hence by 4.4b,

$$\begin{aligned} x\bullet a\bullet A_{(1)}\bullet \{(x_{(1)})^{\dot{j}} ..(x_{(m)})^{\dot{j}}\} &= x\bullet a\bullet A_{(1)}\bullet E^{\{j^m\}}\bullet\{(x_{(1)})^{\dot{j}} .. \ (x_{(m)})^{\dot{j}}\} \\ &= (x\bullet a\bullet A_{(1)}\bullet d)(e\bullet\{(x_{(1)})^{\dot{j}} .. \ (x_{(m)})^{\dot{j}}\}) \end{aligned}$$

for some $d^{\alpha\hat{}}$, $e_{\alpha\hat{}} \in \mathfrak{R}^{+}$ with $de = E^{\{j^m\}}$ (see 3.10a). But the left hand side of this is just $\phi_1(x)$ with $x \in \mathfrak{Q}$. If $f = a\bullet A_{(1)}\bullet d$, and $\xi = e\bullet\{(x_{(1)})^{\dot{j}} .. \ (x_{(m)})^{\dot{j}}\}$, then $f \in \mathfrak{F}^{+}$ and ξ depends only on $\mathfrak{F}$, with $f \in \mathfrak{R}^{+}$ if $\phi_1(x)$ has rational coefficients. This proves the lemma.

4.7 Invariant relations. The notation $\mathfrak{C}_{\beta} \times .. \times \mathfrak{C}_{\varsigma}$ will be used to denote the class of ordered sets of elements, one from each of $\mathfrak{C}_{\beta},..,\mathfrak{C}_{\varsigma}$. Now let $\mathfrak{L}$ be a subclass of $\mathfrak{C}_{\beta} \times .. \times \mathfrak{C}_{\varsigma}$ and let $\phi_i(b,..,\ d,\mathfrak{F})$ $(i = 1,..,\ m)$ be a set of polynomials separately homogeneous in the components of tensors $b_{\beta},..,\ d_{\varsigma}$ in some frame $\mathfrak{F}$. The relations

$$\phi_i(b,..,\ d,\mathfrak{F}) = 0 \qquad (i = 1,..,\ m) \qquad\qquad 4.7a$$

represent an invariant condition on the sets $(b,..,\ d)$ of $\mathfrak{L}$ if this condition on these sets is independent of the frame $\mathfrak{F}$ (see also 7.2). Without loss of generality we can suppose that the degrees $r,..,\ t$ to

which the components of b,.., d respectively appear in ϕ_i are the same for each i. For, any polynomial ϕ of lower degree can be replaced by a set of polynomials each of which is the original ϕ multiplied by suitable powers of the components of b,.., d. Thus, r,..,t may be assumed to be the respective maxima of the degrees to which b,..,d appear in the original polynomials. Conditions 4.7a may then be written as

$$\{(b)^{\wedge}..(d)^{t}\}\cdot f_{(i)} = 0 \qquad (i = 1,.., m)$$

(see 4.3). If 4.7a represents an invariant condition on the elements of $\mathcal{L}$, lemma 4.6 applies where $\mathcal{Q}$ is the set of tensors $\{(b)^{\wedge}.. (d)^{t}\}$ for which $(b,.., d) \in \mathcal{L}$. Then 4.7a may be replaced by

$$\{(b)^{\wedge}.. (d)^{t}\}\cdot F = 0 \qquad 4.7b$$

for elements (b,.., d) of $\mathcal{L}$, where F is some element of $\mathcal{F}$, with $F \in \mathcal{R}$ if the ϕ's have rational coefficients.

This result (together with 4.8) may be used to give a proof of Gram's theorem on invariant equations (see (10) p. 271), to which it is essentially equivalent. If b,.., d are covariant, the tensor F may be regarded as a substitutional operator which may then be expressed as a linear combination of Young tableaux see also (8) p. 117. Hence 4.7b may be expressed as the vanishing of a set of Clebsch forms involving b,.., d.

4.8 Invariants. Let $\mathcal{L}$ be as in 4.7 and $\phi(b,..,d,\mathfrak{F})$ be a polynomial, separately homogeneous in the components of each of b,.., d, in a frame $\mathfrak{F}$, Then if the condition $\phi(b,.., d,\mathfrak{F}) = 0$ is by itself an invariant condition on

the elements of $\mathcal{L}$, lemma 4.6 tells us that

$$\phi(b,..,d) = \zeta\{(b)^{\wedge}..(d)^{t}\}\cdot f$$

for all $(b,..,d) \in \mathcal{L}$, where ζ depends only on $\mathcal{H}$ and where $f^{\beta^{\wedge}..\delta^{t}} \in \mathcal{F}^{+}$ with $f \in \mathcal{H}^{+}$ if ϕ has rational coefficients (r,..,t are the degrees to which b,..,d appear in ϕ as in 4.7). In this case $\phi(b,..,d)$ is an invariant of b,..,d in the Classical sense (see (10) p.169), since if components are taken in a new frame of reference, ϕ is altered by a factor depending only on the equations of the coordinate transformation. The first main theorem of invariant theory (see (10) p.182) follows from the above result since f is an expression involving only I, ϵ and ϵ^{*}. The ϵ and ϵ^{*} give rise to the bracket factors of the Classical invariant formalism. This result is also essentially Littlewood's fundamental theorem for concomitant of a set of tensors (see (6) p.324). The result given here is somewhat more generally stated since the class $\mathcal{L}$ is unrestricted.

5. Tensor rays and geometrical objects.

When invariant theory is used in the study of projective geometry, the geometrical objects considered are generally represented by forms. Such a form can then be represented conveniently by a tensor (see 4.3). Certain geometrical properties of the geometrical objects might then be interpreted in terms of tensor operations concerning the corresponding

tensors. However, tensor addition does not correspond to a geometrically significant operation, since it is really equivalence classes of proportional tensors which represent the geometrical objects and the equivalence class of the sum of two tensors is not determined uniquely by their equivalence classes. In order to obtain the required exact correspondence between geometrical and algebraic operations, the idea of a tensor ray system will be introduced. This deals with the operation of addition in another way (see 5.3).

5.1 Notation. If $a, b \in \mathcal{C}$ and there exists a non-zero element λ of $\mathfrak{R}$ such that $a = \lambda b$, this will be written

$$a \approx b.$$

The relation $\approx$ is clearly an equivalence relation. The equivalence classes of tensors under this relation will be called tensor rays.

Let $\mathcal{S}$ be a subset of $\mathcal{C}$. Then the notation $|\mathcal{S}|'$ will be used to denote the set of tensor rays which are equivalence classes of the elements of $\mathcal{S}$. The notation $|\mathcal{S}|$ will be used to denote the set $|\mathcal{S}|'$ but excluding the equivalence classes of the zeros of $\mathcal{S}$.

I shall usually use the same symbol to denote a tensor ray as to denote one of the tensors belonging to this equivalence class. Thus, in order to avoid ambiguity, the symbol $\approx$ will be used for equality between tensor rays. The notation for pfoducts, contractions, transpositions and the elements $P^{(\sim)}_{(\beta)}$, $N^{(\sim)}_{(\beta)}$, ϵ, ϵ^* of $\mathfrak{R}^+$ will be carried over to the

corresponding tensor rays.

5.2 Tensor ray systems. Let $\mathcal{S} \subseteq \mathcal{C}$ where $\mathcal{S}$ is closed with respect to the operations of outer multiplication and contraction and suppose that $\mathcal{R} \subseteq \mathcal{S}$. Then $|\mathcal{S}|'$ will be called a tensor ray system. Thus a tensor ray system is an algebraic system $|\mathcal{S}|'$ admitting operations of multiplication and contraction and containing $|\mathcal{R}|'$ as a subsystem.

Since $|\mathcal{S}|'$ admits multiplication and contraction it also admits contracted multiplication. In particular, since $\mathcal{R}$ contains all the transposes of each $I_{(\beta)}$, $|\mathcal{S}|'$ is also closed with respect to the operations of transposition. More generally, if $b_\beta, .., c_\gamma \in |\mathcal{S}|'$ and $F_{\beta\ldots\delta}^{\alpha\ldots\varepsilon} \in |\mathcal{R}|'$ where $r, .., t \geqslant 0$, then

$$\{(b)^\alpha .. (c)^\varepsilon\} \bullet F \in |\mathcal{S}|'$$

Conversely, the tensor ray system generated by b,.., c consists of all the elements of the above form. For the elements b, .., c are certainly of this form (choose one of r,.., t to be unity with the rest zero and make F a suitable unit of $\mathcal{C}$). Also the elements of $|\mathcal{R}|'$ are clearly all of this form (choose r = .. = t = 0). Furthermore, $[\{(x)^\alpha .. (z)^\varepsilon\} \bullet F][\{(x)^{\alpha'} .. (z)^{\varepsilon'}\} \bullet F'] = \{(x)^{\alpha+\alpha'} .. (z)^{\varepsilon+\varepsilon'}\} \bullet F''$ where F'' is some transpose of FF'. Finally, any contraction of $\{(x)^\alpha .. (z)^\varepsilon\} \bullet F$ is again of the form $\{(x)^\alpha .. (z)^\varepsilon\} \bullet F'$, where F' is a contraction of F.

5.3 Linear dependence of tensor rays. The operation of addition is not admitted by a tensor ray system but it is possible however to state the more geometrical notion of linear dependence in terms of tensor ray operations. Thus, the condition for $a_{(1)\beta}, \ldots, a_{(\nu)\beta}$ to be linearly dependent is

$$\{a_{(1)} \ldots a_{(\nu)}\} \bullet N^{(\nu)}_{(\beta)} \approx 0.$$

(The components of the tensor $\{a_{(1)} \ldots a_{(\nu)}\} \bullet N^{(\nu)}_{(\beta)}$ are the r^{th} order determinants whose columns are taken from the matrix whose rows are the components of tensors $a_{(1)}, \ldots, a_{(\nu)}$.)

In particular, if $a_\beta, b_\beta \in |\mathcal{C}|$ the case $r = 2$ gives

$$a \approx b \text{ if and only if } (ab) \bullet N^{(2)}_{(\beta)} \approx 0 \qquad 5.3a$$

($a, b \neq 0$ since $|\mathcal{C}|$ does not contain the zeros of $|\mathcal{C}|'$).

5.4 The projective space $\mathfrak{P}$. Each set $|\mathcal{C}_\beta|$ is a projective space over $\mathfrak{R}$. The dimension of $|\mathcal{C}_\beta|$ is $\nu(\beta) - 1$ (see 3.6). In particular, $|\mathcal{C}_{\alpha^*}|$ is a projective $(n-1)$-dimensional projective space over $\mathfrak{R}$. This space will be denoted by $\mathfrak{P}$. $|\mathcal{C}_\alpha|$ is the dual space of $\mathfrak{P}$ and the elements of $|\mathcal{C}_\alpha|$ thus correspond to (and may be identified with) the primes of $\mathfrak{P}$. If $x^\alpha \in |\mathcal{C}_{\alpha^*}|$ and $u_\alpha \in |\mathcal{C}_\alpha|$, then the condition for the point x to lie on the prime u is

$$u \bullet x \approx 0.$$

Other spaces associated with $\mathfrak{P}$ which will have importance here are the many-way product spaces

$$|\mathcal{C}_\beta| \times \ldots \times |\mathcal{C}_\delta|$$

CHAPTER 7

of ordered sets of elements from each of $|\mathcal{C}_\beta|, \ldots, |\mathcal{C}_\delta|$ $(\beta, \ldots, \delta \in \mathfrak{S})$.

5.5 Ray frames. Let $\mathcal{F}$ be a frame for $\mathcal{C}$, the elements of $\mathcal{F}$ being $x^\alpha_{(1)}, \ldots, x^\alpha_{(m)}$; $u_{(1)\alpha}, \ldots, u_{(m)\alpha}$. Define $\bar{\mathcal{F}}$ to be the ordered set $(x_{(0)}, x_{(1)}, \ldots, x_{(m)}$; $u_{(0)}, u_{(1)}, \ldots, u_{(m)})$ where

$$x^\alpha_{(0)} = x_{(1)} + \ldots + x_{(m)} \text{ and } u_{(0)\alpha} = u_{(1)} + \ldots + u_{(m)}.$$

Then the set $|\bar{\mathcal{F}}|$ will be called the ray frame for $|\mathcal{C}|$ corresponding to $\mathcal{F}$.

The points $x_{(1)}, \ldots, x_{(m)}$ of $\mathfrak{P}$ are the vertices of a simplex of reference with $x_{(0)}$ as unit point. The primes $u_{(1)}, \ldots, u_{(m)}$ are the corresponding faces of this simplex of reference and $u_{(0)}$ is the unit prime. Thus a ray frame can be constructed from any set of n + 1 elements $x^\alpha_{(0)}, x^\alpha_{(1)}, \ldots, x^\alpha_{(m)}$ of $|\mathcal{C}_{\alpha^*}|$ no n of which are linearly dependent. Each prime $u_{(i)}$ $(i \neq 0)$ is the join of the points $x_{(1)}, \ldots, x_{(i-1)}, x_{(i+1)}, \ldots, x_{(m)}$. The prime $u_{(0)}$ is the harmonic polar of $x_{(0)}$ with respect to the simplex $(x_{(1)}, \ldots, x_{(m)})$. In terms of ray operations, we have

$$u_{(i)} \approx \epsilon^* \bullet (x_{(1)} \ldots x_{(i-1)} \, I \, x_{(i+1)} \ldots x_{(m)})$$

and

$$u_{(0)} \approx (u_{(1)} \ldots u_{(m)}) \bullet P^{(m)} \bullet \{(x_{(0)})^{m-1} \, I\}$$

as is readily seen on expanding $P^{(m)}$.

5.6 Component ratios. In 4.2, it was shown how the components of a tensor a in a frame $\mathcal{F}$ could be obtained as scalars by means of tensor operations involving a and $\mathcal{F}$. The same process will not work for tensor ray operations since the only two elements of $|\mathcal{C}_1|'$ (scalar tensor rays)

are the equivalence classes of 1 and of 0. However, it is possible to associate the elements of the field $\mathfrak{R}$ with certain elements of $\mathcal{J}$. The condition that the ratio of two components of a in $\mathcal{F}$ should take a certain value can then be stated in terms of tensor ray operations concerning a, $|\bar{\mathcal{F}}|$ and the corresponding element of $|\mathcal{J}|$. The ratios of the components of a tensor in a frame $\mathcal{F}$ will be called the component ratios in $|\bar{\mathcal{F}}|$ of the corresponding tensor ray.

Define

$$R^{[\lambda]\beta_1\beta_2}_{(\beta)\beta_1\beta_2} = I^{(2)\beta_1\beta_2}_{(\beta)\beta_1\beta_2} - \lambda\, I^{(2)\beta_1\beta_2}_{(\beta)\beta_2\beta_1} \qquad 5.6a$$

for each $\beta \in \mathfrak{S}$, $\lambda \in \mathfrak{R}$ and put $R^{[\lambda]}_{(\alpha)} = R^{[\lambda]}$. Then $R^{[\lambda]}_{(\beta)} \in \mathcal{J}$ and $R^{[\lambda]}_{(\beta)} \in \mathcal{R}$ if λ is rational. Provided $n > 1$ and $\beta \neq 1$, the element λ of $\mathfrak{R}$ is uniquely determined by the tensor ray given by $R^{[\lambda]}_{(\beta)}$, since no formally distinct transposes of $I^{(2)}_{(\beta)}$ can be proportional.

Now let a_β be some element of $\mathfrak{C}$. Let $a^{i_1\cdots i_n}_{j_1\cdots j_n}$ be the components of a in $\mathcal{F}$ given by 4.2a. Writing $u_{(i)\alpha} = u'^{\alpha *}_{(i)}$ we have

$$a^{i_1\cdots i_n}_{i_1\cdots i_n} = a\bullet(u'_{(i_1)}\cdots u'_{(i_n)}\; x_{(j_1)}\cdots x_{(j_n)}).$$

Let λ be the ratio of $a^{i_1\cdots i_n}_{j_1\cdots j_n}$ to $a^{k_1\cdots k_n}_{l_1\cdots l_n}$ (where $a^{k_1\cdots k_n}_{l_1\cdots l_n} \neq 0$) so that

$$a\bullet(u'_{(i_1)}\cdots u'_{(i_n)}\, x_{(j_1)}\cdots x_{(j_n)}) - \lambda\, a\bullet(u'_{(k_1)}\cdots u'_{(k_n)}\, x_{(l_1)}\cdots x_{(l_n)}) = 0.$$

Putting $x^{\alpha}_{(0)} = x'_{(0)\alpha *}$ we have $x'_{(0)}\bullet u'_{(i)} = 1 = u_{(0)}\bullet x_{(i)}$ $(i = 1,\cdots n)$.

Hence

$$\{a(x'_{(0)})\hat{}(u_{(0)})^{\Delta}\}\bullet R^{[\lambda]}_{(\beta)}\bullet(u'_{(i_1)}\cdots u'_{(i_n)}\, x_{(j_1)}\cdots x_{(j_n)}\, u'_{(k_1)}\cdots u'_{(k_n)}\, x_{(l_1)}\cdots x_{(l_n)}) \approx 0 \qquad 5.6b$$

and this holds for no other value of λ, so that the fact that λ is this

component ratio can be stated in terms of tensor ray operations concerning a, $|\bar{\mathcal{F}}|$ and $R^{[\lambda]}_{(\beta)}$. Furthermore, the correspondence between $R^{[\lambda]}_{(\beta)}$ and $R^{[\lambda]}$ can be stated entirely in terms of tensor ray properties. For $R^{[\lambda]}$ uniquely determines and is determined by $R^{[\lambda]}_{(\beta)}$, by the property that for each $b^{\alpha}_{\beta} \in |\mathcal{C}_{\alpha^*\beta}|$,

$$R^{[\lambda]} \bullet (b)^2 \approx (b)^2 \bullet R^{[\lambda]}_{(\beta)} \qquad 5.6c$$

Thus, if the elements of $\mathfrak{R}$ are associated with the tensor rays given by the $R^{[\lambda]}$'s, the component ratios of a tensor ray in a frame $|\bar{\mathcal{F}}|$ can be given a tensor ray interpretation.

5.7 Geometrical objects. I shall take as the definition of a geometrical object in $\mathfrak{P}$, the following inductive pair of properties.

(5.7a) Any point of $\mathfrak{P}$ is a geometrical object in $\mathfrak{P}$.

(5.7b) Any algebraic correspondence between geometrical objects in $\mathfrak{P}$ is again a geometrical object in $\mathfrak{P}$.

In 5.7b, the algebraic correspondence may be "between" a single geometrical object, that is it may simply be an algebraic system of geometrical objects. For completeness I shall also allow the vacuous "correspondence between no objects". A zero-dimensional correspondence gives an ordered set of geometrical objects.

This definition of a geometrical object is of considerable generality. It appears to include (or at least to include something exactly corresponding to) anything that one might reasonably call a geometrical object in $\mathfrak{P}$ and which has a finite algebraic definition. (An example of a

non-algebraic geometrical object might be the transformation of a complex projective space induced by the automorphism $z \to \bar{z}$ of the complex field.) Examples of geometrical objects included in this definition are points, primes, linear spaces and algebraic varieties in general, ordered sets of algebraic varieties, birational correspondences, systems of linearly equivalent curves on a surface, etc., etc. A system of geometrical objects which has an infinite number of components would not be included in this scheme as it stands, but if such a system were in some sense finitely based, this would probably not lead to any essential widening of the scheme.

5.8 Tensor ray representation of geometrical objects. I shall now show how any geometrical object in $\mathfrak{P}$ can be represented by an element of $|\mathcal{C}|$ Firstly as in 5.4, any point of $\mathfrak{P}$ can be represented uniquely by ("is") an element of $|\mathcal{C}_{\alpha^*}|$. Secondly, any algebraic correspondence between elements of $|\mathcal{C}_\alpha|, .., |\mathcal{C}_\varsigma|$ can be represented by an algebraic variety in the many-way space $|\mathcal{C}_\alpha| \times .. \times |\mathcal{C}_\varsigma|$. Now this many-way space may be identified with a certain algebraic variety in $|\mathcal{C}_{\alpha..\varsigma}|$, in accordance with the Segre representation of a many-way space in a projective space (see (3) p.93). Thus, if it can be shown that any algebraic variety in $|\mathcal{C}_\eta|$ can be represented by an element of $|\mathcal{C}|$, the representation of geometrical objects defined as by 5.7a and 5.7b will be completed.

Let $\mathcal{U}$ be an algebraic variety in $|\mathcal{C}_\eta|$. Then $\mathcal{U}$ is the union of unmixed varieties $\mathcal{U}_0, \mathcal{U}_1, .., \mathcal{U}_m$ of respective dimensions 0, 1,.., m (some of which

CHAPTER 7

may be of order zero). Now any unmixed algebraic variety may be represented by its Cayley form (see (3) p.32) and any form can be uniquely represented by a certain symmetrized tensor (see 4.3). The usual way of writing a Cayley form is in terms of prime coordinates. It can equally well be written in terms of point coordinates (see (3) p.49 and also (8) p.172). Hence if $\mathcal{U}_\lambda$ is of order s, it can be uniquely represented by an element $f_{(\lambda)\eta^{*}\circ(m-\lambda-1)}$ of $|\mathcal{C}|$ with the property that the condition

$$f_{(\lambda)}\bullet\{(y_{(1)})^{\circ}\,..\,(y_{(m-\lambda-1)})^{\circ}\}\approx 0 \qquad 5.8a$$

is necessary and sufficient for the join of the linearly independent points $y_{(1)}^{\eta^*},\ldots, y_{(m-\lambda-1)}^{\eta^*}$ of $|\mathcal{C}_\eta|$ to meet $\mathcal{U}_\lambda$, where in addition

$$f_{(\lambda)}\bullet(P_{(\eta^*)}^{(\omega)})^{m-\lambda-1}\approx f_{(\lambda)}. \qquad 5.8b$$

(see 4.3). (If $\mathcal{U}_\lambda$ has order zero, we can put $f_{(\lambda)}\approx 1$.)
The variety $\mathcal{U}$ may then be represented by the outer product $f_{(0)}\,f_{(1)}\,..\,f_{(m)}$ of these tensor rays, since the outer product of two elements of $|\mathcal{C}|$ determines the two factors uniquely (see also 5.11b). The above representation agrees with that given for primes in $\mathfrak{P}$ (see 5.4) and, more generally, with the natural representation of a primal of order ε as a symmetric element of $|\mathcal{C}_{\alpha^\varepsilon}|$. It also agrees with the representation of a linear space by its tensor of dual Grassmann coordinates.

In many cases the same tensor ray will represent geometrical objects arising from 5.7a and 5.7b in several quite distinct ways. A trivial

example is the element 1 of $|\mathcal{C}_1|$ which represents a vacuous correspondence or an "algebraic variety" of order zero or of dimension n-1 in $\mathfrak{P}$ (which is itself of dimension n-1). A less trivial example is afforded by a symmetric element $a_{\alpha_1\alpha_2}$ of $|\mathcal{C}_{\alpha^2}|$. This may be interpreted as a quadric primal in $\mathfrak{P}$, the correspondence $a\bullet(xy) \approx 0$ of conjugate points x^α, $y^\alpha \in |\mathcal{C}_\alpha|$, or the system of quadric primal envelopes apolar to it. It is doubtful, however, whether these various interpretations ought to be considered as giving rise to distinct geometrical objects, since the different interpretations are very closely related. In fact they are equivalent in the sense that each determines the other in a geometrical way. It will be shown in 5.12 that this is the case for all the interpretations which can be assigned in the above way to an element of $|\mathcal{C}|$. Thus two interpretations may, if desired, be considered as representing the same geometrical object if they correspond to the same element of $|\mathcal{C}|$.

On the other hand, I shall find it convenient at present to distinguish between closely associated geometrical objects if they correspond to different elements of $|\mathcal{C}|$. (For an alternative point of view see 6.9.) Thus, a linear space would be considered as a different geometrical object, when represented by its tensor ray of Grassmann coordinates (linear system of points) represented by its tensor ray of dual Grassmann coordinates (linear system of primes). I shall thus tend to regard a geometrical object <u>as</u> an element of $|\mathcal{C}|$ together with a suitable geometrical interpretation.

5.9 Canonical interpretation of elements of $|\mathcal{C}|$. Let $a^{\alpha_1 \cdots \alpha_n}_{\alpha_1 \cdots \alpha_s}$ be any element of $|\mathcal{C}|$ and let $y^{\alpha}_{(1)}, \ldots, y^{\alpha}_{(s)}$, $v_{(1)\alpha}, \ldots, v_{(r)\alpha} \in |\mathcal{C}|$. Then the proposition

$$(v_{(1)} \cdots v_{(r)}) \bullet a \bullet (y_{(1)} \cdots y_{(s)}) \approx 0 \qquad 5.9a$$

asserts that a certain multilinear condition holds between the points $y_{(1)}, \ldots, y_{(s)}$ and the primes $v_{(1)}, \ldots, v_{(r)}$. The tensor ray a is uniquely determined by this correspondence between s points and r primes. Thus, every element of $|\mathcal{C}|$ has an interpretation as a geometrical object. I shall call this multilinear correspondence the canonical interpretation of a. This interpretation of an element of $|\mathcal{C}|$ is not a special case of those given in 5.8. (For this to be a special case, some kind of many-way Cayley form would have to have been used in 5.8.) The geometrical equivalence of these various interpretations of elements of $|\mathcal{C}|$ will be proved in 5.12.

5.10 Geometrical operations. In order that the following ideas will apply it will be necessary to assume, henceforth, that $n \geqslant 3$, i.e. that $\mathfrak{P}$ has dimension two or more. (The case $n = 2$ can also be treated if a geometrical operation in $\mathfrak{P}$ is taken to mean one induced on $\mathfrak{P}$ when $\mathfrak{P}$ is embedded in a plane.)

A geometrical property of a set of geometrical objects is one which can be stated ultimately in terms of the notion of collinearity between three points. In particular, the property of an object, that it be a geometrical object as defined by 5.7, is a geometrical one. For, the way

in which the algebra of geometry may be built up from this basic notion (see (11) p.141) implies that the concept of an algebraic correspondence can be stated in terms of the notion of colinearity. Hence, the property of a correspondence that it be the canonical representation of some element of $|\mathcal{C}|$, is a geometrical one.

Now, by 5.3, the condition that three points x^α, y^α, z^α of $\mathfrak{X}$ be colinear may be expressed simply in terms of tensor ray operations, namely

$$(xyz)\cdot N^{(3)} \approx 0. \qquad 5.10a$$

Thus it follows that any geometrical property of a set of geometrical objects can be stated in terms of tensor ray operations, since 5.8a, 5.8b and 5.9a are also stated in terms of such operations.

In practice, however, it would be very cumbersome to have to express every geometrical property in terms of 5.10a. Many geometrical properties can be stated as the vanishing of a set of polynomials with integer coefficients in the components in a frame $\mathfrak{F}$ of a set of tensors, where the polynomials are separately homogeneous in each of the sets of components and where the condition they represent is independent of $\mathfrak{F}$. Thus by 4.7 such a property can be written as

$$\{(b)^\wedge \ldots (d)^t\}\cdot F \approx 0 \qquad 5.10b$$

where $F \in |\mathcal{R}|$, $b_\beta, \ldots, d_\varsigma \in |\mathcal{C}|$.

A property expressible in this form might be called a closed geometrical proposition since the set of points in $|\mathcal{C}_\beta| \times \ldots \times |\mathcal{C}_\varsigma|$ it

represents is a (closed) algebraic variety. I shall call such a variety a geometric variety (see 7.2). A more general type of geometrical proposition would be one corresponding to a geometric variety but with a geometric subvariety removed, a geometric subvariety of this variety replaced, and so on. It can, in fact, be shown that the most general geometrical property is of this form but with enumerable unions of geometric varieties replacing the geometric varieties above.

5.11 Interpretation of tensor ray operations. I shall now show, conversely, that every tensor ray operation has a geometrical interpretation. This can be seen on general grounds but a more direct argument will be given here which leads to an explicit interpretation of most of the concepts involved. In what follows, the geometrical interpretation of each element of $|\mathfrak{C}|$ will be taken to be its canonical interpretation.

(5.11a) Transposition. If $a^{\alpha_1 \dots \alpha_\lambda}_{\alpha_1 \dots \alpha_\lambda} \in |\mathfrak{C}|$, the correspondence between $v_{(1)\alpha}, \dots, v_{(\lambda)\alpha}$; $y^{\alpha}_{(1)}, \dots, y^{\alpha}_{(\lambda)}$ represented by $a^{\alpha_{k(1)} \dots \alpha_{k(\lambda)}}_{\alpha_{l(1)} \dots \alpha_{l(\lambda)}}$ is the correspondence represented by a between these primes and points taken in the orders $v_{(k(1))}, \dots, v_{(k(\lambda))}$; $y_{(l(1))}, \dots, y_{(l(\lambda))}$, given by the respective permutations k and l.

(5.11b) Outer multiplication. If $a^{\alpha_1 \dots \alpha_\lambda}_{\alpha_1 \dots \alpha_\lambda}$, $b^{\alpha_1 \dots \alpha_\rho}_{\alpha_1 \dots \alpha_q} \in |\mathfrak{C}|$ then the condition on $v_{(1)\alpha}, \dots, v_{(\lambda+\rho)\alpha}$; $y^{\alpha}_{(1)}, \dots, y^{\alpha}_{(\lambda+q)}$ given by the canonical representation of ab

$$(v_{(1)} \,..\, v_{(\lambda+\beta)}) \bullet (ab) \bullet (y_{(1)} \,..\, y_{(\alpha+\beta)}) \approx 0$$

is simply the condition that one or other (or both) of

$$(v_{(1)} \,..\, v_{(\lambda)}) \bullet a \bullet (y_{(1)} \,..\, y_{(\alpha)}) \approx 0, \quad (v_{(\lambda+1)} \,..\, v_{(\lambda+\beta)}) \bullet b \bullet (y_{(\alpha+1)} \,..\, y_{(\alpha+\beta)}) \approx 0$$

should hold. Hence the product of two elements of $|\mathcal{C}|$ is geometrically determined by, and also geometrically determines, each of its factors.

(5.11c) Contraction. Let $a^{\alpha_1 \,..\, \alpha_\lambda}_{\alpha_1 \,..\, \alpha_\alpha}$, $v_{(1)\alpha}, \ldots, v_{(\lambda-1)\alpha}$; $y^{\alpha}_{(1)}, \ldots, y^{\alpha}_{(\alpha-1)} \in |\mathcal{C}|$ and put

$$t^{\alpha}_{\alpha} \approx (v_{(1)} \,..\, v_{(\lambda-1)} I) \bullet a \bullet (y_{(1)} \,..\, y_{(\alpha-1)} I).$$

Then the condition that $t^{\bar{\alpha}}_{\bar{\alpha}} \approx 0$ is just the condition that $v_{(1)}, \ldots, v_{(\lambda-1)}$; $y_{(1)}, \ldots, y_{(\alpha-1)}$ should satisfy the correspondence represented by $a^{\alpha_1 \,..\, \alpha_{\lambda-1} \bar{\alpha}}_{\alpha_1 \,..\, \alpha_{\alpha-1} \bar{\alpha}}$. Hence, the operation of contraction has a geometrical interpretation if a geometrical interpretation can be found for the condition $t^{\bar{\alpha}}_{\alpha} \approx 0$, where we may suppose $t \not\approx 0$.

Now suppose that t is such that there exists a simplex for which each vertex $x^{\alpha}_{(i)}$ corresponds in the canonical interpretation of t to the opposite face $u_{(i)\alpha}$. Then we have

$$u_{(i)} \bullet t \bullet x_{(i)} \approx 0 \qquad (i = 1, \ldots, n),$$

so that if a frame is chosen with this simplex as simplex of reference, the matrix of components has all its diagonal elements zero. Thus the trace of this matrix is zero and we have $t^{\bar{\alpha}}_{\bar{\alpha}} \approx 0$.

Conversely, the fact that $t^{\bar{\alpha}}_{\bar{\alpha}} \approx 0$ implies the existence of such a simplex follows from the fact that any square matrix of trace zero, over a

field without characteristic, is equivalent under a similarity transformation to a matrix all of whose diagonal elements are zero. To find such a similarity transformation, we first observe that any (2×2) matrix, which is not a non-zero multiple of the unit matrix, can have its upper left hand element reduced to zero by a similarity transformation. Hence, by repeatedly applying this process to the diagonal (2×2) submatrices of an $(n \times n)$ matrix, and rearranging the rows and columns where necessary, we obtain a matrix with its first n-r diagonal elements zero and a multiple of the $(r \times r)$ unit matrix in its bottom right hand corner. Thus, if the original matrix has trace zero, the reduction of diagonal elements to zero is complete.

If a simplex is found all but one of whose vertices satisfy the required condition with the opposite face, then the relation $t_{\bar{\alpha}}^{\bar{\alpha}} \approx 0$ is necessary and sufficient for the last vertex also to satisfy the condition with its opposite face. This property bears some similarity to the well known condition for apolarity between a quadric primal and a quadric primal envelope. The apolarity condition is, in fact, a direct consequence of this property.

(5.11d) Elements of $|\mathfrak{R}|$. In order to complete the list of geometrical interpretations of tensor ray operations, it is necessary to show that each element of $|\mathfrak{R}|$ has an absolute geometrical interpretation.

Now, consider the elements $R^{[\lambda]}$, defined in 5.6a and let

u_α, v_α, x^α, $y^\alpha \in |\mathcal{C}|$. The condition $(uv)\cdot R^{[\lambda]}\cdot(xy) \approx 0$ can be written in tensor form as $(u\cdot x)(v\cdot y) = \lambda(u\cdot y)(v\cdot x)$. This is the condition that λ should be the cross-ratio that the points x, y make with the primes u, v. Thus, if λ is rational, the condition can be stated in terms of incidence relations involving u, v, x, y and no other fixed elements in $\mathfrak{P}$ (see (11) p.84). That is, $R^{[\lambda]}$ has an absolute geometrical interpretation when λ is rational.

It follows from this that $R^{[\lambda]}_{(\beta)}$ also has an absolute geometrical interpretation whenever λ is rational. For relation 5.6c defines $R^{[\lambda]}_{(\beta)}$ from $R^{[\lambda]}$ by means of the tensor ray operations of multiplication and contraction and an arbitrary element of $|\mathcal{C}_{\alpha^*\beta}|$. Furthermore, 5.6b shows that any element of $|\mathcal{C}|$ with rational component ratios in some ray frame $|\bar{\mathcal{F}}|$ is geometrically definable from $|\mathcal{F}|$. But the elements of $|\mathcal{R}|$ have rational component ratios in all frames. Therefore each element of $|\mathcal{R}|$ has an absolute geometrical interpretation.

This last remark applies also to the elements of $|\mathcal{R}^+|$ (but to no other elements of $|\mathcal{C}|$, see 7.8). Indeed, the canonical interpretations of ϵ and of ϵ^* are the respective conditions that n primes should have a point in common and that n points should lie in a prime. Related to this is the interpretation of $N^{(n)}$, which states that the join of r points meets the intersection of r primes or that the points or primes are linearly dependent. The interpretation of $P^{(n)}$ is that the set of r points

defines an envelope apolar to the set of r primes. Thus the interpretation of $P^{(\lambda)}$ and also of most other elements of $|\mathcal{R}|$ is not so simple to state in terms of incidence relations.

5.12 Equivalence of geometrical interpretations. It remains to prove that the various possible geometrical interpretations given according to 5.8 and 5.9 of an element a of $|\mathcal{C}|$ are in fact geometrically equivalent, that is, each determines the others by its geometrical properties. It will be sufficient to show that the canonical interpretation of a determines each of the others geometrically and that it is the only canonical interpretation of an element of $|\mathcal{C}|$ determining the other interpretations in that particular way, in each case.

Now, consider the inductive definition of a geometrical object and suppose that the required equivalence between the canonical interpretation and that of 5.8 has been established at all stages previous to the one under consideration (the first stage is trivial). Then in relation 5.8a we can suppose that $y_{(1)},\ldots,y_{(n-\lambda-1)}$ are given their canonical interpretations $\mathcal{Y}_1,\ldots,\mathcal{Y}_{n-\lambda-1}$. Let $\mathcal{Z}$ be the canonical interpretation of $f_{(\lambda)}$. Then by 5.11, the relation 5.8a states a geometrical property Θ connecting $\mathcal{Z},\mathcal{Y}_1,\ldots,\mathcal{Y}_{n-\lambda-1}$. But the condition that the point $y_{(1)}$ of $|\mathcal{C}_\eta|$ should lie on $\mathcal{U}_\lambda$ is simply that Θ holds between $\mathcal{Z},\mathcal{Y}_1$ and any arbitrary $\mathcal{Y}_2,\ldots,\mathcal{Y}_{n-\lambda-1}$. Thus $\mathcal{Z}$ geometrically determines $\mathcal{U}_\lambda$ and it follows from 5.11b that the canonical interpretation of $f_{(0)}\ f_{(1)}\ \ldots\ f_{(m)}$

geometrically determines $\mathfrak{U}$. Also this is the only canonical interpretation that determines $\mathfrak{U}$ in this way, provided that 5.8b is satisfied and so also is the restriction on $f_{(\lambda)}$ implied by 5.8a, since $f_{(\lambda)}$ is uniquely determined by these properties and these are tensor ray properties.

6. Geometric functions.

An important question arising out of section 5 is that of deciding which geometrical objects are determined in a geometrical way by a given set of geometrical objects. In view of 5.12 this problem can be reduced to a purely algebraic one concerning tensor ray operations. I shall give here a simple algebraic criterion for characterizing such geometrical objects and some related questions will be discussed. In order to prove the main theorems, I shall borrow the following useful concept from algebraic geometry.

6.1 Generic tensors and tensor rays. Let $\mathfrak{Q}$ be a subclass of the ordered sets of tensors $(x_\beta,.., z_\delta)$ of $\mathfrak{C}_\beta \times .. \times \mathfrak{C}_\delta$ and let $\mathfrak{S}$ be a set of tensors of $\mathfrak{C}$. Then I shall call $(b_\beta,.., d_\delta)$ a generic element of $\mathfrak{Q}$ with respect to $\mathfrak{S}$ if $(b,.., d) \in \mathfrak{Q}$ and if every relation $\phi(b,..,d) = 0$ implies $\phi(x,..,z) = 0$ for all $(x,.., z) \in \mathfrak{Q}$, where $\phi(x,.., z)$ is an expression constructable by means of tensor operations from $x,.., z$ and the elements of $\mathfrak{S}$. A generic element of $|\mathfrak{Q}|$ with respect to $|\mathfrak{S}|$ is

CHAPTER 7

similarly defined but with the ϕ's constructed by means of tensor ray operations. ($|\mathfrak{A}|$ means the class of ordered sets of tensor rays corresponding to the elements of $\mathfrak{A}$.)

This is a generalization of the idea of a generic point of an algebraic variety (see (3) p.10). For if x,.., z are elements of $\mathfrak{R}$ and $\mathfrak{F}$ is a subfield of $\mathfrak{R}$ where $\mathfrak{A}$ is an (irreducible) algebraic variety over $\mathfrak{F}$, then a generic element of $\mathfrak{A}$ with respect to $\mathfrak{F}$ is just a generic point the variety $\mathfrak{A}$, over $\mathfrak{F}$.

An alternative definition of a generic point of an algebraic variety can be given which (unlike the conventional definition) is entirely geometrical in character and does not depend on the frame of reference or the field of definition of the variety. This may be done in terms of tensor rays. For let $\mathfrak{A}$ be an algebraic variety in $\mathfrak{P}$ ($|\mathfrak{A}| \subseteq |\mathfrak{C}_A|$) and let q be the element of $|\mathfrak{C}|$ representing $|\mathfrak{A}|$ according to 5.8. Then if a^α is a generic element of $|\mathfrak{A}|$ with respect to q, it possesses the essential property that is required of a "general" point of $|\mathfrak{A}|$, namely that a lies on no proper subvariety of $|\mathfrak{A}|$ which is geometrically determined by $|\mathfrak{A}|$. The usual definition of a generic point of $|\mathfrak{A}|$, over a field $\mathfrak{Q}$ generated by the component ratios of q in a frame $|\mathfrak{F}|$, would correspond to a generic element of $|\mathfrak{A}|$ with respect to q and $|\mathfrak{F}|$. More generally, a generic point of $\mathfrak{P}$ may be defined with respect to any set $|\mathfrak{S}|$ of geometrical objects in $\mathfrak{P}$ and this definition may be extended to cover any algebraic variety $|\mathfrak{A}|$ in $|\mathfrak{C}_r| \times .. \times |\mathfrak{C}_s|$.

If $|\mathfrak{V}|$ is an absolutely irreducible variety and $|\mathfrak{T}|$ is a finite set of tensor rays then generic elements of $|\mathfrak{V}|$ with respect to $|\mathfrak{T}|$ always exist. This depends essentially on the fact that $\mathfrak{K}$ is of infinite transcendence over the rationals and algebraically closed. It follows from the fact that, in the ordinary sense, a generic point of $|\mathfrak{V}|$, over any finitely generated subfield of $\mathfrak{K}$, exists with coordinates in $\mathfrak{K}$. ($\mathfrak{K}$ may be regarded as a "universal field", see (3) pp.2, 11-14.) Similarly, if $\mathfrak{V}$ is an absolutely irreducible variety and $\mathfrak{T}$ a finite set of tensors, generic elements of $\mathfrak{V}$ with respect to $\mathfrak{T}$ always exist.

6.2 Generic frames. Let $\mathfrak{T}$ be a finite set of tensors $(b_\beta, .., d_\delta)$ and let $\mathfrak{Y}$ be the class of frames for $\mathfrak{C}$. Then I shall call a generic element of $\mathfrak{Y}$ with respect to $\mathfrak{T}$ a generic frame with respect to $|\mathfrak{T}|$ (or with respect to b,.., d). Similarly, a generic ray frame with respect to $|\mathfrak{T}|$ may be defined by replacing $\mathfrak{Y}$ by the class of ray frames for $|\mathfrak{C}|$ (and $\mathfrak{T}$ by $|\mathfrak{T}|$).

Now, the class $\mathfrak{Y}$ of frames for $\mathfrak{C}$ is an absolutely irreducible algebraic variety $\mathfrak{U}$ except for a certain subvariety of $\mathfrak{U}$. For, any frame for $\mathfrak{C}$ is determined (algebraically) by any n linearly independent elements of $\mathfrak{C}$ (see 4.1). It follows, therefore, that generic frames exist with respect to any finite set of tensors. Similarly, generic ray frames always exist with respect to a finite set of tensor rays.

6.3 Rank of a tensor. Let g^{β}_{γ} be any element of $\mathcal{C}_{\beta^*\gamma}$. Then the components of t in any frame $\mathcal{H}$ constitute a matrix G whose rows and columns are given according to the upper and lower indices of g, so that matrix multiplication corresponds to contracted product (see 2.10). If G has rank r, I shall call r the rank of g. This applies also to the tensor ray determined by g. (The definition of rank of course depends on the particular division of indices of g into upper and lowers) All the minors of G of order r+1 vanish while at least one minor of order r does not vanish. Hence, by the determinantal expression for $N^{(\alpha)}_{(\beta)}$ (see 3.7),

$$N^{(\lambda+1)}_{(\beta)} \bullet (g)^{\lambda+1} \approx 0 \text{ and } N^{(\lambda)}_{(\beta)} \bullet (g)^{\lambda} \not\approx 0 \qquad 6.3a$$

6.4 Lemma. Let $|\mathcal{J}|'$ be a tensor ray subsystem of $|\mathcal{C}|'$ and let $g^{\beta}_{\gamma} \in |\mathcal{J}|$. Suppose that the equation $x \bullet g \approx 0$ has a unique solution for $x_{\beta} \in |\mathcal{C}_{\beta}|$, namely $x \approx u_{\beta}$. Then there exists an element v of $|\mathcal{C}|$ such that $v\,u \in |\mathcal{J}|$.

Proof. It follows from the corresponding property for matrices, that the number of linearly independent solutions of the equation $x \bullet g \approx 0$, with $x_{\beta} \in |\mathcal{C}_{\beta}|$, is equal to $\nu(\beta) - r$ where r is the rank of g (see 3.6). Hence $\nu - r = 1$ where $\nu = \nu(\beta)$, so that, by 6.3a, $N^{(\nu-1)}_{(\beta)} \bullet (g)^{\nu-1} \not\approx 0$ and $N^{(\nu)}_{(\beta)} \bullet (g)^{\nu} \approx 0$. But by 3.7c,

$$N^{(\nu)\,\beta_1 \cdots \beta_{\nu-1}\,\bar{\beta}_{\nu}}_{(\beta)\,\bar{\beta}_1 \cdots \bar{\beta}_{\nu-1}\,\bar{\beta}_{\nu}} \; g^{\bar{\beta}_1}_{\gamma_1} \cdots g^{\bar{\beta}_{\nu-1}}_{\gamma_{\nu-1}} \approx N^{(\nu-1)}_{(\beta)} \bullet (g)^{\nu-1} \not\approx 0$$

since $\frac{\nu-(\nu-1)}{\nu-1} \neq 0$, whence

$$0 \not\approx N^{(\nu)}_{(\beta)} \bullet \{(g)^{\nu-1}\ I_{(\beta)}\} \approx h_{\beta}^{\beta_1 \cdots \beta_\nu \gamma_1^* \cdots \gamma_{\nu-1}^*} \quad \epsilon\ |\mathcal{S}|$$

say. We have $h \bullet g \approx 0$, so that for each $w_{\beta^\nu \gamma^{*\nu-1}}$ we have $(w \bullet h) \bullet g \approx 0$ and $w \bullet h \epsilon |\mathcal{C}_\beta|'$. Hence $w \bullet h \approx u$ or $w \bullet h \approx 0$ for every w of $|\mathcal{C}_{\beta^\nu \gamma^{*\nu-1}}|$. It follows from the corresponding matrix property that h has rank 1 and is the outer product of some element $v^{\beta^\nu \gamma^{*\nu-1}}$ with u_β. Thus $vu \epsilon |\mathcal{S}|$ as required.

6.5 Lemma. Let $\mathcal{S}$ be a tensor subsystem of $\mathcal{C}$ and let I, g^β_γ, $h_\gamma \epsilon \mathcal{S}$ Suppose that the equation $x \bullet g = h$ has a unique solution for $x_\beta \epsilon \mathcal{C}_\beta$, namely $x = u$. Then there exists a non-zero element w of $\mathcal{S}$ such that $w u \epsilon \mathcal{S}$

Proof. The number of linearly independent solutions of $x \bullet g = 0$ is zero, since the solution of $x \bullet g = h$ is unique, so that the rank of g is $\nu(\beta) = \nu$. Hence by 3.7e,

$$\left\{ N^{(\nu)}_{(\beta)} \bullet (g)^\nu \right\} u = \sum_{i=1}^{\nu} N^{(\nu)\beta_1 \beta_2 \cdots \cdots \beta_\nu}_{(\beta)\bar{\beta}_1 \cdots \bar{\beta}_{i-1} \beta \bar{\beta}_{i+1} \cdots \bar{\beta}_\nu} \ g^{\bar{\beta}_1}_{\gamma_1} \cdots g^{\bar{\beta}_\nu}_{\gamma_\nu} \ u_{\bar{\beta}_i}$$

In each term of the sum u is contracted with a g^β_γ. This gives h_γ, so the left hand side is an element of $\mathcal{S}$. Thus if $w = N^{(\nu)}_{(\beta)} \bullet (g)^\nu$, we have $w u \epsilon \mathcal{S}$ with $w \epsilon \mathcal{S}$ and $w \neq 0$ (by 6.3a) as required.

6.6 Lemma. Let $\mathcal{S}$ be a tensor system and let a^β_γ, b^β_δ, c_γ, $d_\delta \epsilon \mathcal{S}$. Then there exist elements g^β_ζ, h_ζ of $\mathcal{S}$ such that the equation $x \bullet g = h$ is equivalent to the pair of equations $x \bullet a = c$, $x \bullet b = d$, for elements x_β of $\mathcal{C}_\beta$.

Proof. This result is an extension of 4.5. Put $\zeta = \beta^* \gamma \delta$ and

$$g^{\beta_1}_{\beta_1^* \gamma \delta} = a^{\beta_1}_{\gamma} b^{\beta_2}_{\delta} + a^{\beta_2}_{\gamma} b^{\beta_1}_{\delta} \text{ and } h_{\beta^* \gamma \delta} = c_{\gamma} b^{\beta}_{\delta} + a^{\beta}_{\gamma} d_{\delta}$$

Then the equation $x \cdot g = h$ may be written

$$(x \cdot a)b + a(x \cdot b) = cb + ad \qquad 6.6a$$

This equation is clearly satisfied if $x \cdot a = c$ and $x \cdot b = d$. Conversely suppose x_{β} satisfies 6.6a, which may be rewritten as

$$(x \cdot a - c)b = a(d - x \cdot b). \qquad 6.6b$$

If both sides of this equation are zero then either $x \cdot a = c$ and $x \cdot b = d$ both hold as required, or else one, or both, of a, b is zero. If $a = 0$ and $c \neq 0$ the equations are inconsistent so that the lemma is satisfied with $g = 0$ and $h = c$. If $a = 0$ and $c = 0$ the equation $x \cdot a = c$ is redundant, so we can take $g = b$ and $h = d$. Similarly $b = 0$ can be dealt with. On the other hand, if for some x neither side of 6.6b is zero, choosing a fixed such x satisfying 6.6b and putting

$$p_{\gamma} = x \cdot a - c \text{ and } q_{\zeta} = d - x \cdot b$$

we have $p \neq 0$ and $q \neq 0$, with $pb = aq$. Choosing $y^{\gamma} \in \mathcal{C}$ so that $0 \neq p \cdot y = \lambda$, we have

$$\lambda b = (a \cdot y)q$$

so that putting $r^{\beta} = \frac{1}{\lambda} a \cdot y$ we have $b = r^{\beta} q_{\zeta}$. Hence

$$p(rq) = aq,$$

so that $a = p_{\gamma} r^{\beta} = r^{\beta} p_{\gamma}$, since $q \neq 0$. The equations $x \cdot a = c$ and $x \cdot b = d$ thus become

$$(x \cdot r)p = c \text{ and } (x \cdot r)q = d.$$

These equations are either inconsistent or else one of them is redundant ($x \cdot r$ is a scalar). In either case, the lemma is satisfied with a suitable g and h (as above).

6.7 Theorem 1. Let $|\mathcal{S}|'$ be the tensor ray system generated by the elements $b_\beta, .., d_\delta$ of $|\mathcal{C}|$. Then a necessary and sufficient condition for the element u_ξ of $|\mathcal{C}|$ to be uniquely determined by the geometrical (tensor ray) properties* relating b,.., d and u, is the existence of an element v of $|\mathcal{C}|$ such that

$$v\,u \in |\mathcal{S}|$$

(that is, $uv \approx \{(b)^\wedge .. (d)^t\} \cdot F$ for some $r, .., s \geqslant 0$; $F \in |\mathcal{R}|$.)

Proof. The sufficiency of the condition follows at once from 5.11b and 5.2. For the necessity, let $\mathcal{F}$ be a generic frame with respect to tensors $b', .., d', u'$ corresponding to b,.., d and u (see 6.2). Then if u is geometrically determined by b,.., d, it is certainly geometrically determined by b,.., d and $|\bar{\mathcal{F}}|$. It follows that the component ratios $\xi_1, .., \xi_\ell$ of u in $|\bar{\mathcal{F}}|$ must belong to the field generated by the component

* This is to be taken in a very general sense (see also 5.10). A statement of such a property may involve the use of existential or universal quantifiers, the quantified variables being elements from some $|\mathcal{C}_i|$. Thus, for example, if $\Pi(f,..,h,x)$ is such a property relating f,..,h and x , then the statement "there exists an element x_i of $|\mathcal{C}_i|$ such that $\Pi(f,..,h,x)$ holds" also represents such a property but relating f,..,h. These properties must be ultimately statable in terms of tensor ray operations, tensor ray equality and the logical operations (see (4) p 69).

ratios $\eta_1,..,\eta_m$ of b,.., d in $|\mathfrak{F}|$. For suppose this were not the case. Then we could choose another element $u_{(1)}$ of $|\mathfrak{C}_\xi|$ whose component ratios in $|\mathfrak{F}|$ are conjugates $\xi_1',..,\xi_\ell'$ of $\xi_1,..,\xi_\ell$ over the field $\mathfrak{H}$ generated by $\eta_1,..,\eta_m$. But the tensor ray relations involving b,.., d and u become polynomial relations* connecting $\eta_1,..,\eta_m, \xi_1,..,\xi_\ell$ and these relations will also hold between $\eta_1,..,\eta_m, \xi_1',..,\xi_\ell'$. Thus u would not be distinguished from $u_{(1)}$ by the tensor ray properties connecting it with b,.., d and $|\mathfrak{F}|$. (Any finite extension of $\mathfrak{H}$ can be obtained by using elements of $\mathfrak{R}$, see footnote.)

We have a set of rational functions ψ_i, over the integers, such that

$$\xi_i = \psi_i(\eta_1,..,\eta_m) \qquad (i = 1,.., l)$$

Writing each ξ as a quotient of two components of u′ and each η as a quotient of components of b′,..,or d′ we obtain the relations in the form

$$\phi_i(b',..,d', u') = 0 \qquad (i = 1,.., l) \qquad 6.7a$$

where each ϕ_i is a polynomial with integer coefficients, which is separately homogeneous in the components in $\mathfrak{F}$ of b′,.., d′ and u and linear in the components of u . These relations determine the component ratios of u uniquely. Now, by 4.2a, the ϕ's can be regarded as expressions constructed by means of tensor operations from b′,.., d′, u′

---

* Again this is to be taken in a very general sense (see footnote on p.53). The essential property of such a relation is that it is invariant under any automorphism of $\mathfrak{R}$. It is a consequence of the axiom of choice that an automorphism of $\mathfrak{R}$ will "exist" which sends $\xi_1,..,\xi_\ell$ into any set of conjugates over $\mathfrak{H}$ and leaves $\mathfrak{H}$ invariant, since $\mathfrak{R}$ is of infinite transcendence over $\mathfrak{H}$ (and algebraically closed, without characteristic).

and from $\mathfrak{F}$. But $\mathfrak{F}$ is generic with respect to $b', .., d', u'$. Thus, 6.7a must hold for all choices of frame $\mathfrak{F}$. Hence, by 4.7, the relations 6.7a may be written as

$$\{u(b)^{r'} .. (d)^{t'}\}_{\mathfrak{F}} \cdot A \approx 0$$

where $r', .., t' \geqslant 0$; $A \in |\mathfrak{R}|$ (since all the ϕ's were linear in u). This is a relation of the form $u \cdot g \approx 0$ where

$$g_\theta^\xi \approx \{I_{(\xi)}(b)^{r'} .. (d)^{t'}\}_{\mathfrak{F}} \cdot A \in |\mathcal{S}| .$$

But relations 6.7a determine u uniquely. Thus the equation $x \cdot g \approx 0$ has the unique solution u_ξ if $x \in |\mathcal{C}_\xi|$. The result follows by 6.4.

6.8 Theorem 2. Let $\mathcal{S}$ be the tensor system generated by the elements $b_\beta, .., d_\delta$, I_α^α of $\mathcal{C}$. Then a necessary and sufficient condition for the element u_ξ of $\mathcal{C}$ to be uniquely determined by the tensor properties relating b,.., d and u is the existence of a non-zero element w of $\mathcal{S}$ such that

$$wu \in \mathcal{S}$$

Proof. For the sufficiency we observe simply that if wu and wu′ are both the same element of $\mathcal{S}$, then $w(u - u') = 0$ whence $u' = u$. Thus if $w \in \mathcal{S}$, u is determined by its tensor properties in relation to the elements of $\mathcal{S}$ which are in turn determined by their tensor properties in relation to b,.., d (since I is determined by its tensor properties). For the necessity, a frame $\mathfrak{F}$ generic with respect to b,.., d may be introduced. Then, as in the proof of theorem 1, the components $\xi_1, .., \xi_s$ of u in

belong to a field generated by the components $\eta_1, .., \eta_m$ of b,.., d in $\mathfrak{F}$ That is, there exist polynomials ψ_i, χ_i with integer coefficients such that

$$\xi_i = \frac{\psi_i(\eta_1, .., \eta_m)}{\chi_i(\eta_1, .., \eta_m)} \qquad (i = 1, .., m)$$

Thus if

$$\phi_i(b, .., d, u) = \chi_i(\eta_1, .., \eta_m)\,\xi_i - \psi_i(\eta_1, .., \eta_m) \qquad (i = 1, .., m),$$

the ϕ's are polynomials (with integer coefficients) in the components of b,.., d and u, in $\mathfrak{F}$, which are linear in the components of u and whose simultaneous vanishing uniquely determines the tensor u.

Let $x^{\alpha}_{(1)}, .., x^{\alpha}_{(m)}$; $u_{(1)\alpha}, .., u_{(m)\alpha}$ be the elements of $\mathfrak{F}$. Then by 4.2a the ϕ's may be written

$$\phi_i(b, .., d, u) = \sum_{j} \left\{(u_{(1)})^{\beta_{1ij}} .. (u_{(m)})^{\beta_{mij}}\right\} \bullet k_{(ij)} \bullet \left\{(x_{(1)})^{\gamma_{1ij}} .. (x_{(m)})^{\gamma_{mij}}\right\}$$

where the $k_{(ij)}$ are constructed from b,.., d and u by means of tensor operations. By 4.1c we have

$$\mu\, u_{(i)} = \epsilon^* \bullet (x_{(1)} .. x_{(i-1)} I x_{(i+1)} .. x_{(m)}) \qquad (i = 1, .., m)$$

where $\mu = \epsilon^* \bullet (x_{(1)} .. x_{(m)}) \neq 0$. Hence, for sufficiently large p,

$$\mu^p \phi_i(b, .., d, u) = \sum_{j} (\epsilon^*)^p \bullet t_{(ij)} \bullet \left\{(x_{(1)})^{\lambda_{1ij}} .. (x_{(m)})^{\lambda_{mij}}\right\} \qquad (i = 1, .., m)$$

where each $t_{(ij)}$ is constructed by tensor operations from b,.., d, I and u. Now $\mathfrak{F}$ is a generic frame with respect to b,.., d, u so that each condition $\mu^p \phi_i(b, .., d, u) = 0$ must hold independently of $\mathfrak{F}$. It must hold if $x_{(i)}$ is replaced by $\lambda_i x_{(i)}$ $(\lambda_i \neq 0)$ so that it may be assumed that just one of the $t_{(ij)}$ $(= t_{(i)})$ is non-zero for each i (i.e. each ϕ_i must be isobaric).

It therefore follows from 4.7 that the

conditions $\phi_i = 0$ are equivalent to

$$\{(\epsilon^*)^{\mathfrak{p}} \bullet t_{(i)}\} \bullet F_{(i)} = 0 \qquad (i = 1,.., m),$$

where $F \in \mathcal{R}$, that is, by 3.8a to

$$(N^{(m)})^{\mathfrak{p}} \bullet t_{(i)} \bullet F_{(i)} = 0 \qquad (i = 1,.., m).$$

Now each tensor $t_{(i)}$ is built up from b,.., d, I and u by means of the tensor operations and also is linear in u, since the ϕ's are linear in the components of u. The same remark applies to each tensor $(N^{(m)})^{\mathfrak{p}} \bullet t_{(i)} \bullet F_{(i)}$. Thus we have

$$(N^{(m)})^{\mathfrak{p}} \bullet t_{(i)} \bullet F_{(i)} = u \bullet g_{(i)} - h_{(i)} \qquad (i = 1,.., m)$$

for some $g_{(1)},.., g_{(m)}, h_{(1)},.., h_{(m)} \in \mathcal{S}$, so that the relations $\phi_i = 0$ are equivalent to

$$u \bullet g_{(i)} = h_{(i)} \qquad (i = 1,.., m)$$

and these relations determine u uniquely. But by 6.6, the equations $x \bullet g_{(i)} = h_{(i)}$ may be combined in pairs to form a final equation $x \bullet g = h$; $g, h \in \mathcal{S}$ $(x_\varsigma \in \mathcal{S}_\varsigma)$. This equation has the unique solution $x = u$, so that the result follows from 6.5.

6.9 Geometrically and rationally closed systems. A tensor ray system $|\mathcal{G}|'$ may have the property that for any two elements u, v of $|\mathcal{C}|$, the condition $v\,u \in |\mathcal{G}|$ implies $u \in |\mathcal{G}|$. Such a system will be called geometrically closed. If $|\mathcal{G}|'$ just has the property that for any $u \in |\mathcal{C}|$ and any $w \in |\mathcal{G}|$, the condition $w\,u \in |\mathcal{G}|$ implies $u \in |\mathcal{G}|$, then it will be called rationally closed. In a similar way, geometrically closed

tensor systems and rationally closed tensor systems may be defined. (The elements v and w must, of course, always be taken to be non-zero.) Any geometrically closed system is clearly also rationally closed.

Since any tensor system or tensor ray system contains all the transposes of its elements, the condition that vu (or wu) belongs to the system can be replaced by the condition that some transpose of vu (or wu) belongs to the system and, in particular that uv (or uw) belongs. It follows that for a geometrically closed system, an (outer) product of non-zero elements belongs to the system only if both factors belong to the system.

Theorem 1 tells us that the tensor ray system $|\mathcal{S}|'$ generated by elements b,.., d of $|\mathcal{C}|$ is geometrically closed if and only if every element of $|\mathcal{C}|$ geometrically determined by b,.., d belongs to $|\mathcal{S}|$. Similarly, theorem 2 tells us that the tensor system generated by I and a given finite set of tensors is rationally closed if and only if it contains every tensor which is characterized completely by its tensor properties in relation to the given set of tensors. If $|\mathcal{S}|'$ is any tensor ray system and $|\mathcal{G}|'$ denotes the set of elements u of $|\mathcal{C}|'$ for which v u ϵ $|\mathcal{S}|'$ for some v ϵ $|\mathcal{C}|$, then it is easy to see that $|\mathcal{G}|'$ is a tensor ray system and also that $|\mathcal{G}|'$ is geometrically closed. The system $|\mathcal{G}|'$ is thus the geometric closure of $|\mathcal{S}|'$. In a similar way the geometric closure of a tensor system, and also the rational closure of a system, may be defined. (The geometric closure of a tensor system is not

generally a tensor system, however.) It follows from theorem 1 that the set of elements of $|\mathcal{C}|$ which are geometrically determined by the elements b,.., d of $|\mathcal{C}|$ is just $|\mathcal{G}|$, where $|\mathcal{G}|$ is the geometric closure of the tensor ray system generated by b,.., d.

From some points of view geometrical objects in $\mathfrak{P}$ might be represented more logically not just by single elements of $|\mathcal{C}|$ (see 5.8) but by (finitely generated) geometrically closed tensor ray subsystems of $|\mathcal{C}|$. If this were done, any two geometrical objects which are geometrically equivalent would have the same representation. The condition that one geometrical object determines geometrically another such object would be interpreted as an inclusion relation between the corresponding tensor ray subsystems. However, from the practical point of view, at least, this representation might not be particularly satisfactory.

It is possible to give intrinsic characterizations of geometrically closed systems and of rationally closed systems. (That is, characterizations which do not require reference to elements of $|\mathcal{C}|$ not contained in the system.) For, the condition that $a_{\beta\gamma}$ should factorize into a product $x_\beta y_\gamma$ is simply that $a_{\beta_1\gamma_1} a_{\beta_2\gamma_2} \approx a_{\beta_1\gamma_2} a_{\beta_2\gamma_1}$ (or $a_{\beta_1\bar{\gamma}_1} a_{\beta_2\bar{\gamma}_2} N^{(2)}_{(\gamma)\bar{\gamma}_1\bar{\gamma}_2} \approx 0$). Similarly, the condition on b_γ and $a_{\beta\gamma}$ that $\mathfrak{a} \approx \mathfrak{z}\mathfrak{b}$, for some z_β, may be written $a_{\beta\gamma_1} b_{\gamma_2} \approx a_{\beta\gamma_2} b_{\gamma_1}$ (or $(ab)\bullet(I_{(\beta)} N^{(2)}_{(\gamma)}) \approx 0$).

6.10 Geometric and invariant functions. Let $b_\beta, \ldots, d_\varsigma \in |\mathcal{C}|$. An element a_ς of $|\mathcal{C}|'$ will be called a simple geometric function of b,.., d if it belongs to the tensor ray system $|\mathcal{S}|'$ generated by b,.., d. Thus, the most general form for a simple geometric function is given by

$$a \approx \{(b)^\lambda \ldots (d)^\tau\} \cdot F$$

where $F \in |\mathcal{R}|$. If a is given by a similar expression but with $F \in |\mathcal{P}|$, a will be called a simple invariant function of b,.., d. The simple invariant functions of b,.., d thus constitute the tensor ray system $|\mathcal{I}|'$ generated by b,.., d and the elements of $|\mathcal{P}|$.

A (general) geometric function of b,.., d is an element of $|\mathcal{C}|'$ which belongs to the geometric closure of $|\mathcal{S}|'$. It is thus an element a_ς of $|\mathcal{C}|'$ for which

$$p\, a \approx \{(b)^\lambda \ldots (d)^\tau\} \cdot F$$

for some $p \in |\mathcal{C}|$ and where $F \in |\mathcal{R}|$. Similarly, an invariant function of b,.., d is given if $F \in |\mathcal{P}|$. It is an element of the geometric closure of $|\mathcal{I}|'$.

A rational geometric function of b,.., d is an element of $|\mathcal{C}|'$ which belongs to the rational closure of $|\mathcal{S}|'$. Thus a_ς is a rational geometric function of b,.., d if for some E, $F \in |\mathcal{R}|$,

$$[\{(b)^{\lambda'} \ldots (d)^{\tau'}\} \cdot E]\, a \approx \{(b)^\lambda \ldots (d)^\tau\} \cdot F$$

where $\{(b)^{\lambda'} \ldots (d)^{\tau'}\} \cdot E \not\approx 0$. Similarly, a rational invariant function is given if $E, F \in \mathcal{P}$ and it is thus an element of the rational closure of $|\mathcal{I}|'$.

The terms "simple geometric function", etc., can also be applied to the tensors corresponding to the tensor rays a, b, .., d. The meaning in each case is the same as for the corresponding tensor rays (with the sign $\approx$ indicating proportionality between tensors). By theorem 1, the elements of $|\mathcal{G}|$ which are geometric functions of b,.., d are just those corresponding to geometrical objects geometrically determined by the objects corresponding to b,.., d. The significance of the invariant functions, etc., will appear later (see 8.4, 8.7).

6.11 Concomitants : some examples. In the classical approach to invariant theory, it is the concomitants of a given set of ground forms that are studied. These concomitants are forms which may be considered as invariants of the given ground forms and of a set of variable linear forms (see (10) p.206). If all the forms considered are represented by tensors (in accordance with 4.3), then it may be shown that the concomitants of the ground forms are just those forms whose tensors are linear combinations of expressions built up from the ground form tensors and from ϵ, ϵ^* (and I) by means of the tensor operations of multiplication, contraction and transposition (see (6) p.324 and also 4.8). Thus if b,.., d are the tensors corresponding to the ground forms, the tensors corresponding to the concomitants are just the elements of the tensor system $\mathcal{S}$ generated by b,.., d, ϵ, ϵ^* and the elements of $\mathfrak{R}$. (The element I can be constructed from ϵ and ϵ^*, see 3.8b.) The elements of $\mathcal{S}$

are linear combinations of simple geometric functions of b,.., d, ϵ and ϵ^* of equal weight. Now let $\mathcal{G}$ denote the set of tensors which are geometric functions of b,.., d, so that the tensors belonging to $\mathcal{G}$ are those corresponding to the geometrical objects which are geometrically determined by the objects corresponding to b,.., d. The condition that the concomitants of the given forms should include all those forms which arise from the given forms in a geometrical way is, therefore, that $\mathcal{G} \subseteq \mathcal{S}$.

It will be shown in 7.7 that when the ground forms are "general" forms, it is indeed true that $\mathcal{G} \subseteq \mathcal{S}$. However, it is a very simple matter to construct examples with particular sets of ground forms for which it is not true.

(6.11a) Let there be a single ground form $\phi(x, y)$ which is linear in the two sets of variables (with tensors x^α, y^α) and let ϕ factorize into forms involving x only and y only. Thus $\phi(x, y) \equiv (a \cdot x)(b \cdot y) \equiv (ab) \circ (xy)$ for some a_α, $b_\alpha \in \mathcal{C}_\alpha$. The set $\mathcal{S}$ is thus generated by ϵ, ϵ^*, the tensor ab and the elements of $\mathfrak{R}$. However, $\mathcal{G}$ also contains ab and hence $\mathcal{G}$ contains the two factors a and b. But if n is even, each of ϵ, ϵ^* has an even number of α-indices. This also applies to ab and therefore also to every element of $\mathcal{S}$. But the element a_α of $\mathcal{G}$ has just one α-index and therefore cannot belong to $\mathcal{S}$. The element a is a geometric function of ab but not a simple geometric function of ab, ϵ and ϵ^*.

(6.11b) An objection might be raised to the previous example on the grounds that the tensor ab is reducible in the sense of representation theory. The tensor ab may be expressed as a sum of its symmetric part $(ab)\bullet P^{(2)}$ and its skew-symmetric part $(ab)\bullet N^{(2)}$. If gound forms are chosen corresponding to these tensors (these will be the Clebsch forms $(ab)\bullet(x)^2$ and $(ab)\bullet\{N^{(2)}\bullet(xy)\}$, see 4.4), then the system $\mathcal{S}$ is the same as in 6.11a, but the system $\mathcal{G}$ will not contain either a or b since a and b appear symmetrically as tensor rays in $(ab)\bullet P^{(2)}$ and $(ab)\bullet N^{(2)}$, so that no tensor ray property concerning $(ab)\bullet P^{(2)}$ and $(ab)\bullet N^{(2)}$ can distinguish a from b. It turns out that in this case in fact $\mathcal{G}\subseteq\mathcal{S}$.

However, the question as to whether the ground forms are Clebsch forms is not really relevant to the problem of deciding whether $\mathcal{G}\subseteq\mathcal{S}$ For example, if $n\geqslant 3$ and the single ground form $(ab)\bullet(x)^2$ is chosen (corresponding to the tensor $(ab)\bullet P^{(2)}$), then it will be seen that $(ab)\bullet N^{(2)}\notin\mathcal{S}$ (since $(ab)\bullet N^{(2)}$ changes sign when a and b are interchanged, whereas $(ab)\bullet P^{(2)}$ remains unaltered). However, $(ab)\bullet N^{(2)}$ is certainly a geometric function of $(ab)\bullet P^{(2)}$, because $(ab)\bullet P^{(2)}$ represents a prime pair in $\mathfrak{P}$ and $(ab)\bullet N^{(2)}$ is the tensor of Grassmann coordinates of the intersection of the two primes. In fact, it turns out that the product of $(ab)\bullet N^{(2)}$ with itself is proportional to a transpose of $\{(ab)\bullet P^{(2)}\}^2\bullet(IN^{(2)}I)$. Thus we have $\mathcal{G}\notin\mathcal{S}$ even though the ground form is a Clebsch form.

It would seem that, for the purposes of the present theory, the practice of reducing tensors into irreducible parts provides an unnecessary complication (althought it is naturally important in other contexts). The essential property of factorization, in the case of a tensor such as ab, is to a large extent obscured when only Clebsch forms are considered.

(6.11c) In two of the examples so far considered, the geometric functions of the ground form tensors were not all contained in the systems $\mathcal{S}$, but on the other hand, all the rational geometric functions of the ground form tensors do turn out to be elements of $\mathcal{S}$ in these cases. It is perhaps asking too much of the system $\mathcal{S}$ that it should contain the geometric functions, since they are determined by the original ground forms only up to a constant of proportionality. However, it is no more difficult to construct examples of sets of ground forms for which $\mathcal{S}$ does not even contain all the rational geometric functions arising from these ground forms.

Let a_α, $b_\alpha \in \mathcal{C}_\alpha$ and let the ground forms be a•x and (ab)•(xy), corresponding to the tensors a and ab. Then b is clearly a rational geometric function of a and ab. It is not hard to see that b is not in fact an element of $\mathcal{S}$ (where $n > 1$, of course). Again, it may be objected that (ab)•(xy) is not a Clebsch form. This objection is not serious, since if the ground form (ab)•(xy) is replaced by the Clebsch

form $(ab)\bullet(x)^2$, the system $\mathcal{S}$ is certainly not widened and it can be shown that b is a rational geometric function of a and $(ab)\bullet P^{(2)}$. It is perhaps worth noting that, if n = 2, the forms $a\bullet x$ and $(ab)\bullet(x)^2$ are, by themselves, "general" forms. They are, however, related to one another in a very special way, so that considered together they cannot be regarded as general forms (see 7.7).

(6.11d) The preceding examples may be considered to be artificially constructed to a certain extent, in that the forms considered represent geometrical objects in $\mathfrak{B}$ which are special in some sense, and in fact, reducible. However, the situation is more serious than this, since in the case of Cayley forms, which appear to be the most natural way to represent algebraic varieties by means of forms, the same difficulties seem to arise. If the ground forms are taken to be the Cayley forms of two algebraic varieties, the results obtained in chapter B 4 of (8) indicate that, in general, the Cayley form of the intersection of the two varieties is not a concomitant of the ground forms (except for certain special values of the dimensions and orders).

The main difficulties which arise in the invariant theory of Cayley forms, stem from the fact that the coefficients satisfy complicated non-linear algebraic relations (see (3) p.57 and also (8) p.213). They are not general forms. There are also some corresponding difficulties which arise to a lesser extent even in the case of general forms, where it

often turns out that the system $\mathcal{S}$ is not geometrically closed. This will be shown in 8.7.

The following two lemmas concerning rational geometric functions will be required later.

6.12 Lemma. Let b,.., d $\in \mathcal{C}$ and let a_β be a geometric function of b,.., d which is a linear combination over $\mathfrak{R}$ of simple geometric functions of b,.., d. Then a_β is a rational geometric function of b,.., d.

Proof. Let $\mathcal{S}$ denote the set of elements of $\mathcal{C}$ which are simple geometric functions of b,.., d. Then a is a linear combination of elements $a_{(1)\beta}, \ldots, a_{(\lambda)\beta}$ of $\mathcal{S}$, where we may suppose that $a_{(1)}, \ldots, a_{(\lambda)}$ are linearly independent and that a is not linearly dependent on $a_{(2)}, \ldots, a_{(\lambda)}$. Thus, the two sets $a_{(1)}, \ldots, a_{(\lambda)}$ and a, $a_{(2)}, \ldots, a_{(\lambda)}$ each constitute a basis for the same linear subspace $\mathcal{L}$ of $\mathcal{C}_\beta$. It follows that

$$(a\ a_{(2)} \ldots a_{(\lambda)}) \bullet N^{(\lambda)}_{(\beta)} \approx (a_{(1)} \ldots a_{(\lambda)}) \bullet N^{(\lambda)}_{(\beta)} \qquad 6.12a$$

since each of these two expressions denotes a tensor of Grassmann coordinates of the space $\mathcal{L}$. (This follows from the determinantal expression for $N^{(\lambda)}_{(\beta)}$.) Now, since a is a geometric function of b,.., d, there exists a non-zero element y_γ of $\mathcal{C}$ such that y a $\in \mathcal{S}$. Hence by 6.12a,

$$0 \neq w \approx y\left\{(a_{(1)} \ldots a_{(\lambda)}) \bullet N^{(\lambda)}_{(\beta)}\right\} \in \mathcal{S}$$

But a y $\in \mathcal{S}$, so that aw $\approx$ (ay) $\left\{(a_{(1)} \ldots a_{(\lambda)}) \bullet N^{(\lambda)}_{(\beta)}\right\} \in \mathcal{S}$. Thus, a is a rational geometric function of b,.., d.

As a corollary to this result, we observe that any geometric function

which is also a simple invariant function, must in addition be a rational geometric function (it need not be a simple geometric function). The lemma also implies the more general result that any geometric function which is a linear combination of rational geometric functions is itself a rational geometric function.

6.13 Lemma. Let b,.., d $\in \mathcal{C}$ and let $\lambda_1,.., \lambda_m \in \mathfrak{R}$ and suppose that the element a_β of $\mathcal{C}$ is uniquely determined by the tensor properties relating b,.., d, $\lambda_1,.., \lambda_m$ and a. Then a_β is a linear combination of rational geometric functions of b,.., d.

Proof. If the set of elements of $\mathcal{C}$ which are simple geometric functions of b,.., d is denoted by $\mathcal{S}$ as before, it will be observed that the elements of the tensor system $\mathcal{T}$ generated by b,.., d and the elements of $\mathfrak{R}$ is the set of tensors each of which is a linear combination, over $\mathfrak{R}$, of elements of $\mathcal{S}$. By theorem 2 (6.8), the tensor a satisfies a relation pa = q where $p \neq 0$ and where p, q $\in \mathcal{T}$.

Now, the elements x_β of $\mathcal{C}_\beta$ for which there exists a relation yx = z with $o \neq y \in \mathcal{T}$ and $z \in \mathcal{T}$ constitute a linear subspace $\mathcal{L}$ of $\mathcal{C}_\beta$. For, if yx = z, y'x' = z' and $\mu, \mu' \in \mathfrak{R}$; y', z' $\in \mathcal{T}$, then $(y'y)(\mu x + \mu' x')$ is a linear combination of y'z and a transpose of yz' and is thus an element of $\mathcal{T}$. Furthermore, choosing tensors $x_{(1)\beta},.., x_{(\lambda)\beta}$ which span $\mathcal{L}$ and putting $w_\gamma = y_{(1)} .. y_{(\lambda)}$ $(\neq 0)$ where $y_{(1)},.., y_{(\lambda)}$ are non-zero elements of $\mathcal{T}$ for which $y_{(i)} x_{(i)} \in \mathcal{T}$, (i = 1,.., r), we see

$$wx \in \mathcal{T} \text{ for all } x \in \mathcal{L}.$$

Now the elements w_γ of $\mathcal{T}_\gamma$ which have the above property constitute a linear subspace $\mathcal{M}$ of $\mathcal{T}_\gamma$, since if $wx \in \mathcal{T}$ and $w'x \in \mathcal{T}$, then $(\mu w + \mu' w')\, x \in \mathcal{T}$ $(\mu, \mu' \in \mathfrak{R}\ ;\ w' \in \mathcal{T}_\gamma)$. Since every element of $\mathcal{T}$ is a linear combination of elements of $\mathcal{S}$, there exist elements $g_{(1)}, \ldots, g_{(s)}$ of $\mathcal{S}$ (spanning $\mathcal{T}_\gamma$) such that every element of $\mathcal{M}$ is a linear combination of $g_{(1)}, \ldots, g_{(s)}$. If $t \leq s$, the space spanned by $g_{(1)}, \ldots, g_{(t)}$ meets $\mathcal{M}$ in a subspace whose dimension is less than that of $\mathcal{M}$ by at most $s - t$. Thus, by a suitable choice of t, this subspace of $\mathcal{M}$ can be made to consist of a single ray only. Thus, there is a non-zero element p of $\mathcal{M}$ which is a linear combination of $g_{(1)}, \ldots, g_{(t)}$ but any other element of $\mathcal{M}$ with this property must be proportional to p.

Now by 5.3, the relation of linear dependence can be stated in terms of tensor ray operations. Thus, the sets $|\mathcal{T}|$, $|\mathcal{L}|$ and $|\mathcal{M}|$ can be characterized by their tensor ray properties in relation to the tensor rays of b,.., d. Furthermore, each element of $|\mathcal{S}|$ is completely determined by its tensor ray properties in relation to b,.., d. It follows that the tensor ray of p is also determined by its tensor ray properties in relation to b,.., d. The tensor p is therefore a geometric function of b,.., d. But it is also a linear combination of simple geometric functions of b,.., d, whence, by the previous lemma (6.12) it is a rational geometric function of b,.., d.

Now, the elements v_δ of $\mathcal{T}$ for which $px = v$ for some $x_\beta \in \mathcal{C}_\beta$

constitute a linear subspace $\mathcal{H}$ of $\mathcal{V}_s$ since if $px = v$ and $px' = v'$, then $p(\mu x + \mu' x') = \mu v + \mu' v'$ $(\mu, \mu' \in \mathfrak{R}\ ;\ v' \in \mathcal{V}_s)$. As for $\mathcal{M}$ above, there exist elements $h_{(1)}, .., h_{(u)}$ of $\mathcal{S}$ such that every element of $\mathcal{H}$ is a linear combination of $h_{(1)}, .., h_{(u)}$. Then, as before, if u is chosen suitably, there is an element $q_{(1)}$ of $\mathcal{H}$ which is uniquely determined up to proportionality by the fact that it is a linear combination of $h_{(1)}, .., h_{(u_1)}$ and which therefore must be a rational geometric function of b,.., d (similarly to p above). Renumbering $h_{(1)}, .., h_{(u_1)}$ if necessary, we can replace $h_{(1)}$ by $q_{(1)}$ and every element of $\mathcal{H}$ will be a linear combination of $q_{(1)}, h_{(2)}, .., h_{(u)}$. Then, choosing u suitably, we can find an element $q_{(2)}$ of $\mathcal{H}$ which is uniquely determined up to proportionality by the fact that it is a linear combination of $h_{(2)}, .., h_{(u_2)}$ and which again must be a rational geometric function of b,.., d. Continuing in this manner, we obtain elements $q_{(1)}, .., q_{(k)}$ of $\mathcal{H}$ each of which is a rational geometric function of b,.., d and for which every element of $\mathcal{H}$ is a linear combination of $q_{(1)}, .., q_{(k)}, h_{(k+1)}, .., h_{(u)}$. If the process is carried far enough, there will be no linear combination of $h_{(k+1)}, .., h_{(u)}$ which is an element of $\mathcal{H}$ and therefore $q_{(1)}, .., q_{(k)}$ span $\mathcal{H}$.

The tensor a belongs to the set $\mathcal{L}$ and satisfies the relation $pa = q$ for some $q \in \mathcal{H}$ ($p \in \mathcal{M}$ as above). We have $q = \eta_1 q_{(1)} + .. + \eta_k q_{(k)}$ for some $\eta_1, .., \eta_k \in \mathfrak{R}$ and since $q_{(i)} \in \mathcal{H}$ we have $p\, a_{(i)} = q_{(i)}$ for some $a_{(i)} \in \mathcal{L}$ $(i = 1, .., k)$. Hence

$$p(\eta_1 a_{(1)} + .. + \eta_k a_{(k)}) = q,$$

showing that $a = \eta_1 a_{(1)} + .. + \eta_k a_{(k)}$. But p and $q_{(i)}$ are rational geometric functions of b,.., d so that ep and $f_{(i)} q_{(i)}$ are elements of $\mathcal{S}$ for suitable non-zero e, $f_{(i)} \in \mathcal{S}$. Thus $f_{(i)} ep$ $(\neq 0)$ and $f_{(i)} eq_{(i)}$ are elements of $\mathcal{S}$, so that the relations $(f_{(i)} ep) a_{(i)} = f_{(i)} eq_{(i)}$ imply that each $a_{(i)}$ is a rational geometric function of b,.., d as required.

7. Geometric and invariant varieties

The expression $\{(x)^r..(z)^t\} \cdot F$ with $F \in \mathfrak{M}$ (or $\mathfrak{I}$) defines a simple geometric (or invariant) function of the tensor rays x_α ,.., z_δ for each $(x,..,z) \in |\mathcal{C}_\alpha| \times .. \times |\mathcal{C}_\delta|$. It thus defines a mapping of the many-way space $|\mathcal{C}_\alpha| \times .. \times |\mathcal{C}_\delta|$ into some $|\mathcal{C}_\xi|'$. However, the definitions of the general geometric and rational geometric functions (or invariant functions) of x_β,.., z_δ do not necessarily give rise to a mapping of the whole of $|\mathcal{C}_\beta| \times .. \times |\mathcal{C}_\delta|$, since they depend on the expression $\{(x)^r..(z)^t\} \cdot F$ splitting into factors (see 6.10). This factorization may only occur for sets (x,..,z) belonging to a certain subvariety $|\mathcal{Z}|$ of $|\mathcal{C}_\beta| \times .. \times |\mathcal{C}_\delta|$. Problems concerning the existence and properties of these more general functions would appear to be closely connected with the geometrical properties of such varieties $|\mathcal{Z}|$. That this is so to a certain extent at least, will be shown in (7.5 and 7.6).

7.1 Linear transformation of $\mathcal{C}$. A linear transformation of $\mathcal{C}$ is a mapping of each set $\mathcal{C}_\beta$ onto itself, which preserves the tensor operations and which leaves the elements of $\mathcal{C}_1$ (scalars) invariant. Any linear transformation Φ leaves the α-unit I invariant since I is determined by the property $I \bullet a = a = a \bullet I$ for each $a^\alpha_\alpha \in \mathcal{C}_{\alpha\alpha^*}$. Hence every element of $\mathcal{J}$ is left invariant by Φ.

If $\mathcal{F}$ is any frame for $\mathcal{C}$, then Φ sends $\mathcal{F}$ into another frame $\mathcal{F}'$ since the definition of a frame is given in terms of tensor operations (see 4.2b). The effect of the transformation Φ on any element a of $\mathcal{C}$ is determined by the effect of Φ on the frame $\mathcal{F}$ since the components of a in $\mathcal{F}$ are elements of $\mathcal{C}_1$ obtained from a and $\mathcal{F}$ be means of tensor operations and which are therefore the same as the components of the image a' of a under Φ in the frame $\mathcal{F}'$. Conversely, for any two frames $\mathcal{F}$ and $\mathcal{F}'$, there exists a linear transformation Φ which sends $\mathcal{F}$ into $\mathcal{F}'$ (and Φ is then unique). For, the image of a given element a can be defined as that element of $\mathcal{C}$ whose components in $\mathcal{F}'$ are the same as those of a in $\mathcal{F}$ This is easily seen to determine a linear transformation of $\mathcal{C}$.

A linear transformation Φ of $\mathcal{C}$ induces a collineation of the space $\mathfrak{P}$. If t^α_α is the tensor given by the matrix of the collineation $x' = t \bullet x$, (where x'^α is the image of x^α under Φ,) and if $t^{(-1)\,\alpha}_{\quad\alpha}$ is the tensor given by the inverse of this matrix, then the image $a'^{\alpha_1 \cdots \alpha_\lambda}_{\alpha_1 \cdots \alpha_\lambda}$ under Φ of any element $a^{\alpha_1 \cdots \alpha_\lambda}_{\alpha_1 \cdots \alpha_\lambda}$ of $\mathcal{C}$, is given by

CHAPTER 7

$$a' = (t)^{\wedge} \cdot a \cdot (t^{(-1)})^{\wedge}.$$

A more general type of automorphism Ψ of $\mathcal{C}$ (sending each $\mathcal{C}_\beta$ into itself) may also be defined, which involves an automorphism of the field $\mathcal{C}_1$. The elements of $\mathcal{R}$ are invariant under any such Ψ and the mapping Ψ is determined by its effect on $\mathcal{C}_1$ and a given frame $\mathcal{F}$. Another type of automorphism is a skew automorphism which sends each $\mathcal{C}_\beta$ into the corresponding $\mathcal{C}_\beta^*$.

7.2 Invariant varieties. An algebraic variety $|\mathcal{U}|$ in the many-way space $|\mathcal{C}_\beta| \times \ldots \times |\mathcal{C}_\delta|$ which is invariant under any linear transformation of $\mathcal{C}$, will be called an invariant variety in $|\mathcal{C}_\beta| \times \ldots \times |\mathcal{C}_\delta|$. The equations of $|\mathcal{U}|$ may be written as $\phi_i(x, \ldots, z) = 0$ $(i = 1, \ldots, m)$, where the ϕ's are polynomials separately homogeneous in the components, in a frame $\mathcal{F}$, of each of the tensors $x_\beta, \ldots, z_\delta$. Since $|\mathcal{U}|$ is invariant under any linear transformation, it follows that the equations $\phi_i(x, \ldots, z) = 0$ are together independent of the choice of $\mathcal{F}$ (see 7.1). Thus, by 4.7, the equation of any invariant variety $|\mathcal{U}|$ may be written in the form

$$\{(x)^{\wedge} \ldots (z)^{\wedge}\} \cdot F = 0 \qquad 7.2a$$

where $F \in \mathcal{F}$. If $F \in \mathcal{R}$ (i.e. if the ϕ's may be taken to have rational coefficients), then $|\mathcal{U}|$ will be called a geometric variety in $|\mathcal{C}_\beta| \times \ldots \times |\mathcal{C}_\delta|$. The invariant (or geometric) affine variety $\mathcal{U}$ is the class of sets of tensors $(x, \ldots, z)$ of $\mathcal{C}_\beta \times \ldots \times \mathcal{C}_\delta$ satisfying 7.2a.

An irreducible invariant variety $|\mathcal{U}|$ is an invariant variety which is

not the union of two invariant varieties different from $|\mathcal{U}|$. Similarly, an irreducible geometric variety is one which is not the union of two geometric varieties different from it. An irreducible geometric variety may be reducible when considered as an invariant variety. An irreducible invariant variety is an algebraic variety irreducible over $\mathfrak{K}$, while an irreducible geometric variety need only be irreducible over the rationals (see (8) p. 127). The intersection of any number (or the join of a finite number) of geometric varieties is again a geometric variety, and similarly for invariant varieties.

7.3 Many-way primals. Any hypersurface $|\mathcal{U}|$ (variety of dimension one less than that of the ambiant variety) on $|\mathcal{C}_\alpha| \times \ldots \times |\mathcal{C}_\sigma|$ can be defined by a single equation

$$\{(x)^\lambda \ldots (z)^\kappa\} \cdot f = 0 \qquad 7.3a$$

where $x_\alpha, \ldots, z_\sigma$ are variable elements of $\mathcal{C}$ and where $f^\xi \in \mathcal{C}_{\xi^*}$, $(\xi = \alpha^\lambda \ldots \delta^\kappa)$. (See (3) p. 89 ; 7.3a is a single polynomial equation relating the components of x,.., z in some frame.) Such a variety $|\mathcal{U}|$ will be called a many-way primal in $|\mathcal{C}_\alpha| \times \ldots \times |\mathcal{C}_\sigma|$. It follows from 4.8 that a many-way primal which is also an invariant variety may be given by 7.3a with $f^\xi \in \mathcal{F}^+$. If it is also a geometric variety, we can have $f^\xi \in \mathcal{R}^+$.

Let $|\mathcal{V}|$ be any variety in $|\mathcal{C}_\alpha| \times \ldots \times |\mathcal{C}_\sigma|$ and let $|\mathcal{W}|$ be a hypersurface on $|\mathcal{V}|$ which is the complete intersection of $|\mathcal{V}|$ with a many-way primal $|\mathcal{U}|$ in $|\mathcal{C}_\alpha| \times \ldots \times |\mathcal{C}_\sigma|$. Then if $|\mathcal{U}|$ is given by 7.3a, $|\mathcal{W}|$

is given by those points of $|\mathscr{Q}|$ for which x,.., z satisfy 7.3a. I shall call such a subvariety of $|\mathscr{W}|$ of $|\mathscr{Q}|$ a simple intersection on $|\mathscr{W}|$.

Now suppose that $|\mathscr{W}|$ and $|\mathscr{Q}|$ are both invarient varieties. Then it follows from 4.8 with $\mathscr{L} = \mathscr{Q}$, that for some $g^{\zeta} \in \mathscr{G}^{+}$ (and for some $\zeta \in \mathfrak{R}$)

$$\{(x)^{\wedge}.. (z)^{t}\}\cdot f = \zeta \{(x)^{\wedge}.. (z)^{t}\}\cdot g \quad \text{for all } (x,.., z) \in \mathscr{Q} \qquad 7.3b$$

Thus, any invariant variety, which is a simple intersection on another invariant variety, is the complete intersection of that variety with an invariant many-way primal (given by $\{(x)^{\wedge}.. (z)^{t}\}\cdot g = 0$). There is also the corresponding result for geometric varieties, with $g^{\zeta} \in \mathscr{R}^{+}$.

7.4 The varieties $\mathscr{Q}\{b,.., d\}$ and $\mathscr{Q}\{b,.., d ; \mathfrak{R}\}$. Let $b_{\alpha},.., d_{\delta}$ be non-zero elements of $\mathscr{C}$ (or the corresponding elements of $|\mathscr{C}|$). Then the geometric (affine) variety $\mathscr{Q}\{b,.., d\}$ will be defined as the class of all sets of tensors $(x_{\alpha},.., z_{\delta})$ which satisfy every relation

$$\{(x)^{\wedge}.. (z)^{t}\}\cdot A = 0 \qquad 7.4a$$

where $r,.., t \geqslant 0$ and $A \in \mathscr{R}$, for which

$$\{(b)^{r}.. (d)^{t}\}\cdot A = 0.$$

$|\mathscr{Q}\{b,.., d\}|$ is the corresponding variety in $|\mathscr{C}_{\alpha}| \times .. \times |\mathscr{C}_{\delta}|$. This is clearly a geometric variety since it is the intersection of the geometric varieties given by 7.4a. The variety $|\mathscr{Q}\{b,..,d\}|$ may be regarded as the class of specializations in a tensor ray sense of the set of tensor rays (b,.., d) (regarding b,.., d as tensor rays). The set of tensor rays (b,.., d) is a generic element of $|\mathscr{Q}\{b,.., d\}|$ in the sense of 6.1.

The invariant varieties $\mathscr{V}\{b,..,d;\mathfrak{A}\}$ and $|\mathscr{V}\{b,..,d;\mathfrak{A}\}|$ are defined similarly, but where the A's are allowed to be elements of $\mathfrak{F}$. The set of tensor rays (b,.., d) is a generic element $|\mathscr{V}\{b,..,d;\mathfrak{A}\}|$ with respect to $\mathfrak{F}$ (see 6.1).

Every variety $|\mathscr{V}\{b,..,d\}|$ is an irreducible geometric variety and every $|\mathscr{V}\{b,..,d;\mathfrak{A}\}|$ is an irreducible invariant variety (see (8) p.132). Furthermore, if $|\mathscr{U}|$ is any irreducible geometric variety in $|\mathscr{C}_\beta|\times..\times|\mathscr{C}_\delta|$, then $|\mathscr{U}| = |\mathscr{V}\{b,..,d\}|$ for some $b_\beta \in \mathscr{C}_\beta,.., d_\delta \in \mathscr{C}_\delta$. (This follows from the fact that every algebraic variety which is irreducible (and defined) over a finitely generated subfield of $\mathfrak{A}$ has a generic point with coordinates in $\mathfrak{A}$, see also 6.1.) The corresponding property for irreducible invariant varieties does not hold, however (see (8) p.148).

If $b_\beta,.., d_\delta$ are elements of $|\mathscr{C}|$ and a_ξ is a geometric function of b,.., d given by

$$a\ p \approx \{(b)^\lambda .. (d)^t\}\bullet F$$

($p_\eta \not\approx 0$, $F \in |\mathscr{R}|$), then this defines a geometric function for every $(x,.., z)\in|\mathscr{V}\{b,..,d\}|$. For, the condition for $\{(x)^\lambda .. (z)^t\}\bullet F$ to factorize in the required way is that x,.., z satisfy a certain condition of the form $\{(x)^\lambda .. (z)^t (x)^\lambda .. (z)^t\}\bullet(F)^2\bullet M \approx 0$ which is also satisfied by b,.., d ($M\in|\mathscr{R}|$ and is a transpose of $N^{(2)}_{(\xi)} I^{(2)}_{(\eta)}$, see 6.9). Thus, the geometric variety giving the field of definition of this geometric function, contains the geometric variety $|\mathscr{V}\{b,..,d\}|$. The same remark also applies

to any rational geometric function of b,.., d. For the invariant functions of b,.., d this applies instead, to the variety $|\mathscr{V}\{b,..,d\ ;\mathfrak{R}\}|$.

7.5 Theorem 3. Let $b_\beta,..,\ d_\delta$ be non-zero elements of $\mathscr{C}$ and suppose that every hypersurface on $|\mathscr{V}\{b,..,\ d\}|$ is a simple intersection on $|\mathscr{V}\{b,..,\ d\}|$. Then the geometric functions of b,.., d are just the simple geometric function of b,.., d, ϵ, ϵ^*. If instead, every hypersurface on $|\mathscr{V}\{b,..,\ d\ ;\mathfrak{R}\}|$ is a simple intersection, then every invariant function of b,.., d is a simple invariant function of b,.., d, ϵ, ϵ^*.

Proof. Consider the case of a geometric function $a_\eta \neq 0$ of b,.., d given by

$$p_\theta a_\eta = \{(b)^\wedge..\ (d)^\tau\}\bullet F \qquad 7.5a$$

where $p_\theta \neq 0$ and $F^\xi_{\theta\eta} \in \mathscr{R}$ ($\beta^\wedge..\delta^\tau = \xi$). (The case a = 0 is trivial.) We have

$$\{(b)^\wedge..\ (d)^\tau\}^2\bullet A = 0$$

where $A^{\xi_1\xi_2}_{\theta_1\eta_1\theta_2\eta_2} = F^{\xi_1}_{\theta_1\eta_1} F^{\xi_2}_{\theta_2\eta_2} - F^{\xi_1}_{\theta_1\eta_2} F^{\xi_2}_{\theta_2\eta_1}$ since $p_{\theta_1} a_{\eta_1} p_{\theta_2} a_{\eta_2} = p_{\theta_1} a_{\eta_2} p_{\theta_2} a_{\eta_1}$. Hence, $\{(x)^\wedge..\ (z)^\tau\}\bullet A = 0$ for all $(x,..,\ z) \in \mathscr{V}\{b,..,\ d\}$ (see 7.4). Thus if g^θ, h^θ, u^η, $v^\eta \in \mathscr{C}$, we have

$$\begin{aligned}&[\{(x)^\wedge..(z)^\tau\}\bullet F\bullet(gu)]\ [\{(x)^\wedge..(z)^\tau\}\bullet F\bullet(hv)]\\ &= [\{(x)^\wedge..(z)^\tau\}\bullet F\bullet(gv)]\ [\{(x)^\wedge..(z)^\tau\}\bullet F\bullet(hu)]\end{aligned} \qquad 7.5b$$

for all $(x,..,z) \in \mathscr{V}\{b,..,\ d\}$. For suitable choices of g, h, u and v, these expressions will not vanish for all $(x,..,z) \in \mathscr{V}\{b,..,\ d\}$, since in particular, we can have $(p\bullet g)(a\bullet u)(p\bullet h)(a\bullet v) \neq 0$. Thus, equating

the left-hand side of 7.5b to zero, we get the equation of a hypersurface on $|\mathscr{P}\{b,..,d\}|$ which consists of two component hypersurfaces $|\mathscr{U}_u^g|$ and $|\mathscr{U}_v^h|$, depending respectively, on g, u and on h, v. Since the right-hand side of 7.5b will then also be zero, this hypersurface will also consist of the two corresponding hypersurfaces $|\mathscr{U}_v^g|$ and $|\mathscr{U}_u^h|$. (The hypersurface $|\mathscr{U}_u^g|$ is given by $\{(x)^\lambda..(z)^\tau\}\cdot F\cdot(gu) = 0$, etc.)

Now, for a generic choice of the set $(g^\theta, h^\theta, u^\eta, v^\eta)$, the hypersurfaces $|\mathscr{U}_u^g|$ and $|\mathscr{U}_v^h|$ will have a common component $|\mathscr{W}|$ (possibly null) which is a hypersurface on $|\mathscr{P}\{b,..,d\}|$ common to all the $|\mathscr{U}^:|$'s and which is also the common component of $|\mathscr{U}_v^g|$ and $|\mathscr{U}_u^h|$. The hypersurfaces $|\mathscr{U}_u^g|$ and $|\mathscr{U}_v^g|$ will also have a common component $|\mathscr{W}^g|$ (possibly null), other than $|\mathscr{W}|$, which is a hypersurface on $|\mathscr{P}\{b,..,d\}|$ common to all the $|\mathscr{U}^g|$'s and which thus depends only on g. Similarly, the hypersurface $|\mathscr{W}_u|$ may be defined as the common component of $|\mathscr{U}_u^g|$ and $|\mathscr{U}_u^h|$, which thus depends only on u. But the sum of $|\mathscr{U}_u^g|$ and $|\mathscr{U}_v^h|$ is the same as the sum of $|\mathscr{U}_v^g|$ and $|\mathscr{U}_u^h|$. Therefore $|\mathscr{U}_u^g|$ is the union of its intersections with $|\mathscr{U}_v^g|$ and $|\mathscr{U}_u^h|$ and it follows that $|\mathscr{U}_u^g|$ is the sum of $|\mathscr{W}^g|$, $|\mathscr{W}_u|$ and $|\mathscr{W}|$. If it is assumed that every hypersurface on $|\mathscr{P}\{b,..,d\}|$ is a simple intersection, then we have

$$\{(x)^\lambda..(z)^\tau\}\cdot F\cdot(gu) = [\{(x)^{\lambda_2}..(z)^{\tau_2}\}\cdot f_{(g)}]\,[\{(x)^{\lambda_1}..(z)^{\tau_1}\}\cdot w_{(\omega)}] \qquad 7.5c$$

for all $(x,..,z)\epsilon\,\mathscr{P}\{b,..,d\}$, where $\{(x)^{\lambda_1}..(z)^{\tau_1}\}\cdot w_{(\omega)} = 0$ gives the

equation of $|\mathcal{W}_u|$ and $\{(x)^{\lambda_1}..(z)^{t_2}\}\bullet f_{(g)} = 0$ gives the equation of the sum of $|\mathcal{W}^g|$ and $|\mathcal{W}|$.

Let $u^{\eta}_{(1)}, .., u^{\eta}_{(l)}$ be a basis for $\mathcal{C}_{\eta^*}$ and let $u^{*}_{(1)\eta}, .., u^{*}_{(l)\eta}$ be the corresponding dual basis. That is,

$$u^{*}_{(i)}\bullet u_{(j)} = \delta_{ij} \quad \text{and} \quad \sum_{i=1}^{l} u_{(i)} u^{*}_{(i)} = I_{(\eta)}$$

(see also 4.1a, 4.1b). Let $g^{\theta}_{(1)}, .., g^{\theta}_{(m)}$ and $g^{*}_{(1)\theta}, .., g^{*}_{(m)\theta}$ be similarly defined with respect to $\mathcal{C}_{\theta^*}$, so that

$$g^{*}_{(i)}\, g_{(i)} = \delta_{ij} \quad \text{and} \quad \sum_{i=1}^{m} g_{(i)}\, g^{*}_{(i)} = I_{(\theta)}\,.$$

Then we have

$$F = F\bullet(I_{(\theta)}I_{(\eta)}) = \sum_{i=1}^{m}\sum_{j=1}^{l} F\bullet(g_{(i)}g^{*}_{(i)}u_{(j)}u^{*}_{(j)}),$$

whence 7.5c becomes

$$\{(x)^{\wedge}..(z)^{t}\}\bullet F = \left[\{(x)^{\lambda_1}..(z)^{t_2}\}\bullet f\right]\left[\{(x)^{\lambda_1}..(z)^{t_2}\}\bullet w\right] \qquad 7.5d$$

for all $(x,.., z)\in \mathcal{V}\{b,.., d\}$, where

$$f^{\phi}_{\theta} = \sum_{i=1}^{m} f_{(g_{(i)})}g^{*}_{(i)} \quad \text{and} \quad w^{\xi}_{\eta} = \sum_{j=1}^{l} w_{(u_{(j)})}u^{*}_{(j)}$$

Now as $(x,..,z)$ varies throughout $\mathcal{V}\{b,.., d\}$ and u varies throughout $\mathcal{C}_{\eta^*}$, the equation

$$\{(x)^{\wedge}..(z)^{t}\}\bullet F\bullet(I_{(\theta)}u) = 0$$

defines a geometric subvariety of $|\mathcal{V}\{b,.., d\}| \times |\mathcal{C}_{\eta^*}|$. Also, this subvariety is a hypersurface which is a simple intersection on $|\mathcal{V}\{b,.., d\}| \times |\mathcal{C}_{\eta^*}|$ since by 7.5d it is the same as that given by

$$\{(x)^{\lambda_1}.. (z)^{t_2}\}\bullet w\bullet u = 0.$$

Hence by 7.3b, we may suppose without loss of generality that $w\in \mathcal{R}^{+}$.

But the elements of $\mathscr{R}^+$ are just the simple geometric functions of ϵ, ϵ^* (see 3.9a). Also, comparing 7.5d with 7.5a, we see that

$$a \approx \{(b)^{\wedge} .. (d)^{\epsilon}\}.w.$$

Thus a is a simple geometric function of b,.., d, ϵ, ϵ^* as required.

The proof for invariant functions is exactly similar but with $\mathscr{S}$ replacing $\mathscr{R}$, etc.

7.6 Remarks. It would be desirable to be able to state a converse result to theorem 3, and it does appear that something might be done along these lines owing to the close relationship between (systems of) hypersurfaces on $|\mathscr{U}\{b,.., d\}|$ and geometric functions of b,.., d. However, the converse of theorem 3 would not be true as it stands, as an example will show. Let $x_{(0)}^{\alpha}$, $x_{(1)}^{\alpha}$, .., $x_{(n)}^{\alpha}$ be n + 1 elements of $|\mathscr{C}_{\alpha^*}|$ no n of which are linearly dependent (i.e. the vertices of a ray frame) and let $b \approx x_{(0)} x_{(1)} .. x_{(n)}$. Then the varieties $|\mathscr{U}\{b\}|$ and $|\mathscr{U}\{b ; \mathfrak{A}\}|$ are the same and are given by the product of n + 1 projective spaces each of dimension n - 1. It is clear that not every hypersurface on $|\mathscr{U}\{b\}|$ is a complete intersection with a primal in $|\mathscr{C}_{\alpha^{*n+1}}|$. But every geometric (or invariant) function of b turns out to be a simple geometric (or invariant) function of b and ϵ^*. (In particular, we have $x_{(n)} \approx (\epsilon^* I).b.$) The trouble seems to arise from the fact that $|\mathscr{U}\{b\}|$ possesses invariant hypersurfaces which are not simple intersections. An adequate converse to theorem 3 would have to take these invariant hypersurfaces into account.

A similar difficulty arises in the statement of an analogous result to theorem 3 which gives a sufficient condition for every rational geometric function of b,.., d to be a simple geometric function of b,.., d, ϵ, ϵ^*. The condition is, firstly, that every hypersurface on $|\mathfrak{P}\{b,.., d\}|$, which can be made into a simple intersection by adding a simple intersection to it, is itself a simple intersection and, secondly, that every geometric hypersurface on $|\mathfrak{P}\{b,.., d\}|$ is a simple intersection. In the case of a variety $|\mathfrak{P}\{b\}|$, part of this condition is implied by the condition that $|\mathfrak{P}\{b\}|$ be normal.

7.7 General forms. A set of ground forms will be called general forms if their corresponding tensor rays b,.., d ($\epsilon|\mathfrak{C}|$) have the property that $|\mathfrak{P}\{b,.., d\}|$ is a many-way linear space. Thus, b,.., d may satisfy linear symmetry conditions, but apart from this, they must be "generic" as regards their geometrical properties. Theorem 3 implies that the tensor ray system generated by b,.., d, ϵ and ϵ^* is geometrically closed since any hypersurface on a many-way linear space is a simple intersection (see (3), p...). It follows from this (see 6.11) that the concomitants of a set of general ground forms do, in fact, include all those forms which are geometrically determined by these ground forms (but see also 8.7).

Other types of ground forms whose concomitants have this property are given by the forms representing general linear sets. For example, if $a_{(1)}{}^{\bullet}(x)^{\wedge}, .., a_{(s)}{}^{\bullet}(x)^{\wedge}$ are general forms representing s primals of

order r in $\mathfrak{S}_\beta$ ($a_{(i)\alpha^\wedge}$, $x^\alpha \in \mathfrak{C}$), then the form $(a_{(1)} \dots a_{(\omega)}) \cdot N^{(\omega)}_{(\alpha^\wedge)} \cdot \{(x_{(1)})^\wedge \dots (x_{(\omega)})^\wedge\}$ represents the pencil of primals they determine. The fact that the concomitants of this form include all the forms it geometrically determines, follows from theorem 3 and Severi's result that every hypersurface on a Grassmann variety is a complete intersection with a primal (see (3) p.368), since the variety $|\mathfrak{V}\{b\}|$ arising here is a Grassmann variety.

7.8 Subsystems of $|\mathfrak{F}^+|$. If b,.., d are elements of $|\mathfrak{R}^+|$, then $|\mathfrak{V}\{b,.., d\}|$ consists of a single point only. Hence, theorem 3 implies that the system $|\mathfrak{R}^+|'$ is geometrically closed and is thus the geometric closure of $|\mathfrak{R}|'$. (This can also be shown directly using 4.6.) Thus, the elements of $|\mathfrak{C}|$ which have an absolute geometrical significance are just the elements of the system $|\mathfrak{R}^+|$ (see 5.11d).

If b,.., d are any elements of $|\mathfrak{F}^+|$, then $|\mathfrak{V}\{b,.., d, \mathfrak{R}\}|$ is a single point. Hence $|\mathfrak{F}^+|'$ is also geometrically closed and is the geometric closure of $|\mathfrak{F}|'$. It follows from this that the tensor ray system $|\mathfrak{G}|'$ generated by the elements b,.., d, ϵ, ϵ^* of $|\mathfrak{F}^+|$ is also geometrically closed. For suppose p a $\epsilon |\mathfrak{G}|$. Then we must have p $\epsilon |\mathfrak{F}^+|'$. But it can be shown that any non-zero element of $\mathfrak{F}$ can be contracted to give a non-zero scalar (see (8) p. 47). Thus, if p is multiplied by a suitable power of ϵ or ϵ^* and suitably contracted, an element ≈ 1 can be produced. Hence a $\epsilon |\mathfrak{G}|$.

Considerations of weight (see 3.9) show that if b,.., d $\epsilon |\mathcal{J}|$, then the tensor ray system $|\mathcal{J}|'$ generated by b,.., d is rationally closed. For, the elements of $|\mathcal{J}|$ are the elements of $|\mathcal{G}|$ which have weight zero. Thus if p a $\epsilon |\mathcal{J}|$ and p $\epsilon |\mathcal{J}|$ then a also must have weight zero and so a $\epsilon |\mathcal{J}|$ since a $\epsilon |\mathcal{G}|$. In particular, $|\mathcal{R}|$ is rationally closed and, similarly, so also is $|\mathcal{P}|$

One consequence of this is that if $R^{[\lambda]}$ is defined as in 5.6a for $\lambda \epsilon \mathfrak{R}$, we have that $R^{[\lambda+\mu]}$, $R^{[\lambda-\mu]}$, $R^{[\lambda\mu]}$, $R^{[\lambda/\mu]}$ and $R^{[\lambda]}_{(\beta)}$ are all simple geometric functions of $R^{[\lambda]}$ and $R^{[\mu]}$ ($\mu \neq 0$). Hence, the elements $R^{[\lambda]}$ which belong to a tensor ray system $|\mathcal{T}|'$ are those for which λ belongs to a certain field determined by $|\mathcal{T}|$. This is a field of absolute invariants of the generators of $|\mathcal{T}|'$. The most general absolute invariants determined by the generators of $|\mathcal{T}|'$ are the elements of the corresponding field determined by the geometric closure of $|\mathcal{T}|'$, since these are the cross-ratios geometrically determined by the generators of $|\mathcal{T}|$ (see 5.11d). (This field is also that given by the rational closure of $|\mathcal{T}|$.) The elements of any tensor ray subsystem of $|\mathcal{P}|$ constitute a system like $|\mathcal{P}|$ (see 3.3) but with the field $\mathfrak{R}$ replaced by a subfield of $\mathfrak{R}$.

8. Groups of linear transformations.

In the projective geometry of a space $\mathfrak{P}$, properties which are invariant under the group of all projective transformations of $\mathfrak{P}$ are

studied. Certain other geometries are obtained if this group is replaced by one of its subgroups (see Klein's Erlanger Program (5)). I shall be concerned here with groups of such transformations which leave a certain (finite) set of tensors or tensor rays invariant. The question as to which other tensors or tensor rays are left invariant under such a group of transformations will be answered in 8.4 and 8.5.

8.1 Lemma. Let $|\mathcal{S}|'$ and $|\mathcal{T}|'$ be two isomorphic, finitely generated tensor ray subsystems of $|\mathcal{C}|'$ and let a be any element of $|\mathcal{C}|'$. Then there exists an element b of $|\mathcal{C}|'$ with the property that the given isomorphism between $|\mathcal{S}|'$ and $|\mathcal{T}|'$ can be extended to an isomorphism between the tensor ray system generated by a and $|\mathcal{S}|'$, and that generated by b and $|\mathcal{T}|'$, with a and b as corresponding elements.

The exactly analogous result for tensor subsystems of $\mathcal{C}$ also holds.

Proof. Let $|\bar{\mathcal{F}}|$ be a ray frame for $|\mathcal{C}|'$ which is generic with respect to all the generators of $|\mathcal{S}|'$ and of $|\mathcal{T}|'$ (see 6.2). Let $\xi_1, \ldots, \xi_m$ be the component ratios of the generators of $|\mathcal{S}|'$ in $|\bar{\mathcal{F}}|$ and let $\eta_1, \ldots, \eta_m$ be the corresponding component ratios of the corresponding generators of $|\mathcal{T}|'$ in $|\bar{\mathcal{F}}|$. Now any polynomial relation $\phi(\xi_1, \ldots, \xi_m) = 0$ (where ϕ has integer coefficients) can be written as the vanishing of an expression constructed from the elements of $|\mathcal{S}|'$ and of $|\bar{\mathcal{F}}|$, by means of tensor ray operations. Conversely, the condition that such a tensor ray expression should vanish can be written as the vanishing of a set of such polynomials ϕ. The

corresponding result also holds for $|\mathfrak{J}|'$ and the η's. But the tensor ray system generated by $|\mathfrak{I}|'$ and $|\overline{\mathfrak{F}}|$ is isomorphic with that generated by $|\mathfrak{J}|'$ and $|\overline{\mathfrak{F}}|$, since $|\mathfrak{I}|'$ and $|\mathfrak{J}|'$ are isomorphic and $|\overline{\mathfrak{F}}|$ is a generic frame with respect to $|\mathfrak{I}|'$ and $|\mathfrak{J}|'$ (see 6.1). Hence, the fields generated by $\xi_1, .., \xi_m$ and by $\eta_1, .., \eta_m$ are isomorphic, with ξ_i corresponding to η_i ($i = 1, .., m$). Let $\xi_{m+1}, .., \xi_l$ be the component ratios of a in $|\overline{\mathfrak{F}}|$ Then elements $\eta_{m+1}, .., \eta_l$ of $\mathfrak{R}$ can be chosen so that this isomorphism extends to an isomorphism between the fields generated by $\xi_1, .., \xi_l$ and by $\eta_1, .., \eta_l$, with ξ_i corresponding to η_i ($i = 1, .., l$) (see (3) p. 2). (We may choose $\eta_{m+1}, .., \eta_l$ in $\mathfrak{R}$, since $\mathfrak{R}$ is of infinite transcendence over these fields and algebraically closed.) Now choose the element b of $|\mathfrak{C}|$ to have the component ratios $\eta_{m+1}, .., \eta_l$. Then the isomorphism between the fields generated by $\xi_1, .., \xi_l$ and by $\eta_1, .., \eta_l$ implies a corresponding isomorphism between the tensor ray system generated by a, $|\mathfrak{I}|'$ and $|\overline{\mathfrak{F}}|$, and that generated by b, $|\mathfrak{J}|'$ and $|\overline{\mathfrak{F}}|$. This establishes the isomorphism required for the lemma.

The proof for tensor systems is exactly analogous.

8.2 Theorem 4. Let $b_\beta, .., d_\delta$, $b'_\beta, .., d'_\delta \in |\mathfrak{C}|$, let $|\mathfrak{I}|'$ be the tensor ray system generated by b, .., d and the elements of $|\mathfrak{J}|$ and let $|\mathfrak{J}|'$ be the tensor ray system generated by b', .., d' and the elements of $|\mathfrak{J}|$. Then a necessary and sufficient condition for the existence of a linear transformation of $\mathfrak{C}$ which sends b, .., d into b', .., d' is that the systems

$|\mathcal{J}|'$ and $|\mathcal{J}|'$ should be isomorphic, with the elements of $|\mathcal{F}|$ self-corresponding and with b,.., d corresponding to b',.., d'.

The exactly analogous result, for which $b_\alpha,\ldots, d_\delta$ are tensors and for which $\mathcal{J}$ and $\mathcal{J}$ are tensor systems, also holds (with $\mathcal{J}$, $\mathcal{J}$, $\mathcal{F}$ replacing $|\mathcal{J}|'$, $|\mathcal{J}|'$, $|\mathcal{F}|$, respectively).

Proof. Consider the case of tensor rays. The condition is obviously necessary since the elements of $|\mathcal{F}|$ are invariant under all linear transformations of $\mathcal{C}$ (see 7.1). For the sufficiency, let $|\bar{\mathcal{F}}|$ be any ray frame for $|\mathcal{C}|$ and let $\xi_1,\ldots,\xi_m$ be the component ratios of b,.., d in $|\bar{\mathcal{F}}|$. Now the elements $\xi_1,\ldots,\xi_m$ of $\mathfrak{A}$ may each be associated with one of the elements $R^{[\xi_1]}_{(\zeta)},\ldots, R^{[\xi_m]}_{(\theta)}$ of $|\mathcal{F}|$ (see 5.6a), and the fact that $\xi_1,\ldots,\xi_m$ are the component ratios of b,.., d in $|\bar{\mathcal{F}}|$, can then be stated as the vanishing of a set of expressions constructed from $R^{[\xi_1]}_{(\zeta)},\ldots, R^{[\xi_m]}_{(\theta)}$, b,.., d and the elements of $|\bar{\mathcal{F}}|$ be means of tensor ray operations (see 5.6b). Now, let $|\mathcal{J}_0|'$ and $|\mathcal{J}_0|'$ be, respectively, the tensor ray system generated by $R^{[\xi_1]}_{(\zeta)},\ldots, R^{[\xi_m]}_{(\theta)}$, b,.., d and that generated by $R^{[\xi_1]}_{(\zeta)},\ldots, R^{[\xi_m]}_{(\theta)}$, b',.., d'. The systems $|\mathcal{J}_0|'$ and $|\mathcal{J}_0|'$ are, respectively, subsystems of $|\mathcal{J}|'$ and $|\mathcal{J}|'$, and hence they are isomorphic, with the elements $R^{[\xi_1]}_{(\zeta)},\ldots, R^{[\xi_m]}_{(\theta)}$ self-corresponding. By 8.1, the elements of a new frame $|\bar{\mathcal{F}}_1|$ may be introduced in turn, with the property that the system generated by $|\mathcal{J}_0|'$ and $|\bar{\mathcal{F}}|$ is isomorphic with that generated by $|\mathcal{J}_0|'$ and $|\bar{\mathcal{F}}_1|$. The nature of this isomorphism implies that the component ratios of b,.., d

in $|\bar{\mathcal{F}}|$ are the same as the corresponding component ratios of b,.., d in $|\bar{\mathcal{F}}_1|$. Hence, if the linear transformation of $\mathcal{C}$ is chosen so as to send the frame $|\bar{\mathcal{F}}|$ into $|\bar{\mathcal{F}}_1|$, it must also send b,.., d into b′,.., d′ as required (see 7.1).

The proof for tensors b,.., d′ is similar.

8.3 Absolute invariants. If b,.., d, b′,.., d′ $\epsilon\, |\mathcal{C}|$, another way of stating the condition given in theorem 4 for the existence of a linear transformation sending b,.., d into b′,.., d′ is

$$\mathcal{A}\{b,..,\ d\ ;\mathfrak{A}\} = \mathcal{A}\{b',..,\ d';\mathfrak{A}\}$$

It is not sufficient that the geometrical properties of b,.., d should be the same as those of b′,.., d′ (i.e. $\mathcal{A}\{b,..,\ d\} = \mathcal{A}\{b',..,\ d'\}$) and that the absolute invariants (defined in the sense of invariant theory as the quotient of two relative invariants of equal weight) of forms corresponding to b,.., d and to b′,.., d′, should be the same in each case. For let n = 3 and $x^{\alpha}_{(1)},..,\ x^{\alpha}_{(4)}$; $y^{\alpha}_{(1)},..,\ y^{\alpha}_{(4)}$ represent two sets of four collinear points in $\mathfrak{P}$. Now there will exist a linear transformation which sends $x_{(1)},..,\ x_{(4)}$ into $y_{(1)},..,\ y_{(4)}$ only if the cross-ratio of the four points is the same in each case. But the forms $u\bullet x_{(1)},..,\ u\bullet x_{(4)}$ (forms in u_α) clearly have no non-zero relative invariants (e.g. $\epsilon^{*}\bullet(x_{(1)}\, x_{(2)} x_{(3)}) = 0$) and hence they have no absolute invariants (in the sense of invariant theory).

To cover such a situation, a more general definition of an absolute

invariant is necessary. Such a definition was suggested at the end of 7.8 as the set of cross-ratios (or elements $R^{[\lambda]}$) which are geometrically determined by b,.., d. It appears that this definition is in fact adequate for the above purpose. The variety $|\mathcal{V}\{b,..,d\ ;\mathfrak{A}\}|$ may be represented by an element v of $|\mathcal{C}|$ in accordance with 5.8. Since this variety is an invariant variety, it will follow that v must be an element of $|\mathcal{J}^+|$ (see 8.8). The component ratios of v are therefore the same in any frame and hence the $R^{[\lambda]}$'s corresponding to these component ratios must be geometric functions of b,.., d. These component ratios determine $\mathcal{V}\{b,..,\ d\ ;\mathfrak{A}\}$ completely.

groups of transformations? that $\exists$ an element $\epsilon|\mathcal{C}|$ whose invariance characterizes the group ..... cayley tensor .....

8.4 Theorem 5. Let $b_\beta,..,\ d_\delta \epsilon |\mathcal{C}|$. Then a necessary and sufficient condition that the element a_ζ of $|\mathcal{C}|$ be left invariant by every linear transformation of $\mathcal{C}$ which leaves b,.., d invariant is that a be an invariant function of b,.., d. (That is, p a $\approx\{(b)^\lambda..(d)^\tau\}\cdot F$ for some $p\epsilon|\mathcal{C}|$, $F\epsilon|\mathcal{J}|$.)

Proof. The condition is clearly sufficient since the elements of $|\mathcal{J}|$ are invariant under all linear transformations of $\mathcal{C}$ and the product pa determines the tensor ray a uniquely ($p \neq 0$). Conversely, if a_ζ is left invariant by all the linear transformations of $\mathcal{C}$ which leave b,.., d invariant, then, by theorem 4, there is no element a'_ζ of $|\mathcal{C}_\zeta|$ with the property that the tensor ray system generated by b,.., d, a' and the elements of $|\mathcal{J}|$ is isomorphic with that generated by b,.., d, a and the

elements of $|\mathcal{J}|$, where b,.., d and the elements of $|\mathcal{J}|$ are invariant under the isomorphism and where a corresponds to a . That is, a is uniquely determined (in $|\mathcal{C}_S|$) by the relations it satisfies, of the form

$$\{(a)^q (b)^r .. (d)^t\} \cdot A \approx 0$$

with $A \in |\mathcal{J}|$. Now, by the Hilbert basis theorem, a finite number of these relations will suffice to determine a uniquely, say those given when A is each of $A_{(1)}, .., A_{(m)}$ in turn. Thus, a is uniquely determined by the tensor ray properties relating b,.., d, $A_{(1)}, .., A_{(m)}$ and a, so that by theorem 1 (6.7), a is a geometric function of b,.., d, $A_{(1)}, .., A_{(m)}$. Hence a is an invariant function of b,.., d (see 6.10).

8.5 Theorem 6. Let b,.., d $\in \mathcal{C}$. Then a necessary and sufficient condition that the tensor a be left invariant by every linear transformation of $\mathcal{C}$ which leaves b,.., d invariant is that it be a linear combination of rational geometric functions of b,.., d.

Proof. The condition is clearly sufficient since the simple geometric function of b,.., d are certainly left invariant by all the linear transformations of $\mathcal{C}$ which leave b,.., d invariant, (the elements of $\mathcal{R}$ are invariant) and hence the rational geometric functions of b,.., d are also left invariant by these transformations. For the converse, it follows from theorem 4 (by an argument analogous to that given in the proof of theorem 5) that a is uniquely determined by the relations it satisfies, of the form

$$\chi(a, b, .., d) = 0,$$

where the χ's are constructed from a, b,.., d and the elements of by means of tensor operations. As before, a finite number of these relations is sufficient to determine a uniquely. The result then follows from 6.13, where $\lambda_1,.., \lambda_m$ are the elements of $\mathfrak{A}$ arising in these relations (finite in number).

8.6 Littlewood's conjecture. It was conjectured by Littlewood ((7) p. 388) that any form which is left invariant by the group of linear transformations which leave a set of ground forms invariant is necessarily a concomitant of the ground forms. However, we saw in 6.11c that for certain sets of ground forms, the rational geometric functions of the ground form tensors are sometimes tensors corresponding to forms which are not concomitants of the ground forms. Theorem 6 implies that in these cases, there will exist forms which are not concomitants of the ground forms but which, nevertheless, are left invariant under the group. One example (6.11c) of such a set of ground forms is given by a binary quadratic form together with a binary linear form which is a factor of the quadratic. The other factor is then not a concomitant of the ground forms but is invariant under any transformation which leaves the ground forms invariant. Another example is given by a single cubic ground form $(aab)\cdot(x)^3$ with a repeated factor (a_κ, b_κ, x^α, $y^\alpha \in \mathfrak{C}$). The form $(a\cdot x)^2(b\cdot y)$ is then left invariant by the group, but it is not a

concomitant.

However, Littlewood's conjecture is true for the case of a set of general ground forms (see 7.7). (It is possible that the conjecture was only intended to apply to ground forms with indeterminate coefficients.) This result follows from the fact that the rational geometric functions of the ground form tensors b,.., d (being geometric functions of b,.., d) are all simple geometric functions of b,.., d, ϵ and ϵ^* when the ground forms are general forms (see 7.7). Thus by theorem 6, the tensors left invariant by the linear transformations which leave b,.., d invariant are linear combinations of simple geometric functions of b,.., d, ϵ, ϵ^* and are, thus, tensors corresponding to concomitants of the ground forms (see 6.11). It should be observed that some of the tensors which correspond to concomitants are only relatively invariant under the group. This case can also be covered by theorem 6, simply by adjoining ϵ (or ϵ^*) to the ground form tensors. For then every linear transformation of the group must have unit determinant, and the relative invariants all become absolute invariants. Alternatively, each concomitant tensor may be multiplied by a suitable power of ϵ or ϵ^* so that it becomes absolutely invariant under the group (weight zero). (Different linear combinations would be allowed in the two cases, however.)

8.7 Invariant functions of general forms. It is interesting to note for the case of a set of general ground forms that, although the forms left absolutely invariant by the linear transformations which leave the ground forms invariant are all concomitants of these ground forms (see 8.6), the same need not be true of those forms which are altered by a constant of proportionality depending only on the transformation.

A simple example to illustrate this fact is afforded by the case of a single ground form which is a binary sextic. It is clear that the only projectivity on a line which leaves a set of six general points invariant is the identity. Therefore every linear form is altered only by a constant depending on the transformation, when a transformation which leaves the ground form invariant is applied. However, such a linear form cannot be a concomitant since a binary sextic possesses no linear covariants (see (2) p.156). On the other hand, it is true that the square of every linear form is a concomitant of the sextic. This is implied by the results of 8.6, since the linear form can at most change sign when the transformation is applied. (There are only two possible transformation matrices, namely the identity matrix and its negative.)

This example shows that the set of tensors corresponding to the

concomitants of a binary sextic is not geometrically closed. By theorem 5 it is invariant functions of the ground form tensors that are relevant here. These invariant functions need not always be simple invariant functions, even in the case of general ground forms.

8.8 Invariant tensors and tensor rays. A special case of 8.6 is given when the set of ground forms is vacuous, or consists solely of the absolute form $u \cdot x$ (u_α, $x^\alpha \in \mathcal{C}$) corresponding to the tensor I. The result in this case shows that the elements of $\mathcal{I}$ are the only tensors left absolutely invariant under all linear transformations of $\mathcal{C}$. The corresponding result for tensor rays is that the elements of $|\mathcal{I}^+|$ are just those elements of $|\mathcal{C}|$ which are left invariant by all linear transformations of $\mathcal{C}$. This follows from theorem 5 and the fact that $|\mathcal{I}^+|$ is the geometric closure of $|\mathcal{I}|$ (see 7.8).

These results may also be proved directly using 4.6.

8.9 Geometrical algebras. We have seen that the tensor ray system $|\mathcal{C}'|$ provides an algebraic structure which may be used to describe the geometry of the projective space $\mathfrak{P}$. The algebraic properties of elements of $|\mathcal{C}|$ and the geometrical properties of geometrical objects in $\mathfrak{P}$ correspond to one another exactly. The question that now arises is whether such "geometrical algebras" can be constructed which treat geometries other than just the projective geometry over $\mathfrak{R}$.

For many cases of geometries which can be obtained from the projective geometry by restricting the group of transformations under which the geometrical properties are preserved (see (12) p.14), this can indeed be easily done. Such a geometry can be described by the same system $|\mathcal{C}|'$ as before, but with the basic roles of the subsystems $|\mathcal{R}^+|$ and $|\mathcal{F}^+|'$ taken over by systems $|\mathcal{G}|'$ and $|\mathcal{H}|'$, respectively, where $|\mathcal{H}|'$ consists of all tensor rays left invariant by all the transformations of the group, and where $|\mathcal{G}|'$ is a geometrically closed subsystem of $|\mathcal{H}|'$ whose elements correspond to those geometrical objects which have an absolute geometrical significance in the new geometry and the geometrical operations may be defined in relation to $|\mathcal{G}|$. Such concepts as geometric and invariant varieties can then be redefined in relation to $|\mathcal{G}|$ and $|\mathcal{H}|$.

As an example, consider the case of Euclidean three-dimensional geometry. Here n = 4, and the system $|\mathcal{G}|'$ contains a symmetric tensor ray $g^{\alpha\lambda}$ corresponding to the absolute conic, where $(u)^2 \cdot g \approx 0$ is its tangential equation $(u_\alpha \in |\mathcal{C}_\alpha|)$. $|\mathcal{G}|'$ is the set of geometric functions of g and is generated by g, ϵ, ϵ^* and p, where p_α represents the absolute plane. (It is not a simple geometric function of g.) Among the elements of $|\mathcal{G}|$ is the one representing the system of all spheres. (This set of spheres will then constitute a geometric variety for the new geometry.) Another such element represents the system of all pairs of lines which

intersect in a **given rational angle.** Among the elements of $|\mathcal{H}|$ are those which represent the system of all spheres of a given radius and the system of all pairs of lines which intersect in a given **general** angle. These sets will constitute invariant varieties for the geometry. The system $|\mathcal{H}|'$ contains both $|\mathcal{G}|$ and $|\mathcal{F}|$ but is not generated by their elements. It also contains elements which depend on the fact that the distance between two points is an absolute invariant under the group of Euclidean transformations. Thus, $|\mathcal{H}|$ contains elements $d_{\alpha\bullet}$ with the property that $d\bullet\{(x)^2(y)^2\} \approx 0$ $(x^\alpha, y^\alpha \epsilon |\mathcal{C}|)$ if and only if the points x and y are a certain fixed distance apart. (d is a linear combination of $(p)^4$ and an expression constructed from $(g)^2$ and $(\epsilon)^2$.) Points at infinity are excluded in Euclidean geometry, but the corresponding elements of $|\mathcal{C}_{\alpha^*}|$ must be retained to represent the systems of parallel lines.

Another geometry which can be treated in this way is the non-Euclidean geometry obtained by replacing the absolute conic by a proper quadric. Representing this quadric by the symmetric tensor ray h_{α^2} (where $h\bullet(x)^2 \approx 0$ is its point equation), we obtain $|\mathcal{G}|'$ as the set of geometric functions of h. But the form $h\bullet(x)^2$ is in this case a general form (see 7.7). Hence $|\mathcal{G}|'$ is the tensor ray system generated by h. Also, the system $|\mathcal{H}|'$ turns out, in this case, to be the system generated by h and the elements of $|\mathcal{F}|$. The group of transformations arising here is (apart from reality considerations) the orthogonal group (see (12) p.65). Similarly if h is replaced by a general skew-symmetric tensor ray $S_{\alpha 2}$ (n even),

the symplectic group is obtained (see(12) p165).

Affine geometry may be treated in the same way. The systems $|\mathcal{G}|'$ and $|\mathcal{H}|'$ are generated, respectively, by p_∞ (the prime at infinity) and by p_∞ and the elements of $|\mathcal{F}|$. ($p \bullet x$ is a general form.) If a geometry is constructed similarly but with $|\mathcal{G}|'$ generated by p_∞ and q^∞ (where $p \bullet q \neq 0$) and with $|\mathcal{H}|'$ generated by p, q and $|\mathcal{F}|$, then an affine geometry relative to the fixed point q is obtained. The interest of this geometry lies in the fact that it is the one that corresponds to the ordinary vector (and tensor) algebra. Thus, the operation of addition of vectors is geometrically significant and the zero vector has an absolute interpretation. This geometry is, therefore, equally well described by a tensor system as by the tensor ray system $|\mathcal{C}|'$ taken relative to p and q. Indeed, it is possible to embed the tensor system in $|\mathcal{C}|$ in such a way that the tensor operations of addition, multiplication, contraction and transposition can all be interpreted as tensor ray operations in $|\mathcal{C}|'$, involving p and q. In this way, many of the results concerning tensor systems may be deduced from the corresponding results for tensor ray systems. In particular, it is possible to deduce theorem 2 from theorem 1.

I have been concerned here with geometries over a field $\mathfrak{K}$, where $\mathfrak{K}$ is algebraically closed and without characteristic. However, many geometries are defined over fields which are not algebraically closed,

In particular, Euclidean geometry should strictly be defined over the real field. If $\mathfrak{k}$ is not taken to be algebraically closed, some of the results given here do not remain true without modification. Theorem 1, for instance, is not true over the real field. A cubic primal which consists of three primes one of which is real, the other two being conjugate imaginary, determines the real prime geometrically (the other two are not primes in the same sense), but it is clear that this prime cannot be a geometric function of the cubic in the sense of 6.10, since otherwise each of the other two primes would also have to be geometric functions of the cubic. The results of section 5 remain true, however. The status of the assumption that $\mathfrak{k}$ be without characteristic is somewhat less certain. The definitions of $P^{(\wedge)}$ and $N^{(\wedge)}$ given in 3.5 depend on this assumption, so that the theory would have to be completely reorganized in order to deal with fields of finite characteristic.

It is an interesting question whether a geometrical algebra might be constructed to describe such geometries as birational geometry and Riemannian geometry. The groups of transformations in these two cases are not groups of linear transformations, so that a somewhat different approach would be necessary. Perhaps it is not unreasonable to expect to be able to treat Riemannian geometry in this way, since it was in this connection that the tensor calculus was originally discovered. Also, it is perhaps not too fanciful to suggest that a more adequate

algebra for the description of the physical space might be sought along these lines. The non-existence of an actual physical origin or frame of reference is a strong indication that a calculus which does not describe space in terms of such a frame of reference might be valuable. Physical objects could perhaps be identified with certain elements of an algebraic structure. The result of making a measurement (or of performing an experiment in physics) might then be interpreted somewhat similarly to the way components of a tensor may be obtained by contracting it with the elements of a frame $\mathscr{F}$. The status of a frame $\mathscr{F}$ in the geometry is somewhat similar to that of an observer together with his measuring apparatus.

References

(1) Eisenhart, L. P. Riemannian Geometry (Princeton) : 1926.

(2) Grace, J. H. & Young, A. Algebra of Invariants (Cambridge) : 1903.

(3) Hodge, W. V. D. & Pedoe, D. Methods of Algebraic Geometry, vol.2 (Cambridge) : 1952.

(4) Kleene, S. C. Introduction to Metamathematics (Amsterdam) : 1952.

(5) Klein, F. Gesamm. math. Abh. p.409 : 1921.

(6) Littlewood, D. E. Invariant theory, tensors and group characters. Phil. Trans. A, No.807, 239, 305 : 1944.

(7) Littlewood, D. E. On invariant theory under restricted groups. Phil. Trans. A, No.809, 239, 387 : 1944.

(8) Penrose, R. Tensor Methods in Algebraic Geometry. Ph.D. Dissertation, Cambridge, 1956. (unpublished)

(9) Rutherford, D. E. Substitutional Analysis (Edinburgh) : 1948.

(10) Turnbull, H. W. The Theory of Determinants Matrices and Invariants, 2nd Ed. (Glasgow) : 1945.

(11) Veblen, O. & Young, J. W. Projective Geometry, vol.1 (Boston) : 1910.

(12) Weyl, H. Classical Groups (Princeton) : 1946.

8
A SELF-REPRODUCING ANALOGUE

Nature **179** (1957), 1183. With L. S. Penrose

The explanatory diagram that appears on p.353 was added and sent out with the reprints, but it did not actually appear in the article as published by *Nature*.

LETTERS TO THE EDITORS

The Editors do not hold themselves responsible for opinions expressed by their correspondents. No notice is taken of anonymous communications.

A Self-reproducing Analogue

THE most striking peculiarity of living organisms is their property of self-reproduction. The most elementary forms, virus or phage particles, can reproduce themselves in favourable circumstances only, and this principle applies also to the multiplication of nucleic acid complexes in chromosomes. It is sometimes thought that the self-reproducing properties of nucleic acid depend upon its highly complex structure. Consequently, any mechanical analogue for self-reproduction would involve very intricate mechanisms. This does not seem to be so, and the device described here has the critical reproductive property although it is of the simplest character.

A flat material, such as plywood or vulcanite, is cut into pieces with shapes corresponding to *A* or *B* drawn to scale in Fig. 1. These objects are placed on a track where they can slide freely though they cannot pass one another. The track, formed by a shelf or groove, is blocked at both ends and restricted by a roof. Horizontal shaking will cause the pieces to move both on account of their own inertia and friction and on account of collisions with one another and with the ends of the track. The pieces will not link up in these circumstances.

Two pieces, *A* and *B*, hooked together, as shown in Fig. 2, are now added to those originally on the track, and shaking is resumed. The result is to reproduce the same *AB* complex at any point on the track where an *A*-piece happens to be immediately on the left of a *B*-piece. If the experiment is repeated, with the alteration that the new pieces inserted are hooked together as *B* and *A* in Fig. 3, the result will be to reproduce this figure, *BA*, at all possible places along the track.

Table 1

(i) Original neutral arrangement.
A.B.B.A.B.A.A.A.B.A.B.B.
(ii) Result after adding the complex *AB*.
AB... AB.B.AB.A.A.AB.AB.B.
(iii) Result after adding the complex *BA*.
BA... A.B.BA.BA.A.A.BA.B.B.

Thus, if the pieces are arranged as in Table 1, addition of *AB* produces four new *AB* groups and addition of *BA* produces three new *BA* groups. The difference between *AB* and *BA* is analogous to a mutation in that the changed complex is repeated in its changed form. To emphasize the analogy, asymmetrical markings are shown on the surfaces of the pieces. The forms chosen here are not the only ones suitable for this type of demonstration. They need not be symmetrical, though this is useful as it increases stability in the neutral position (Fig. 1).

The example given here shows how reproduction can be demonstrated by an exceedingly simple mechanism. It has been the starting point for construction of more complicated models with similar basic properties.

L. S. PENROSE

University College,
Gower Street,
London, W.C.1.

R. PENROSE

Bedford College,
Regent's Park,
London, N.W.1.

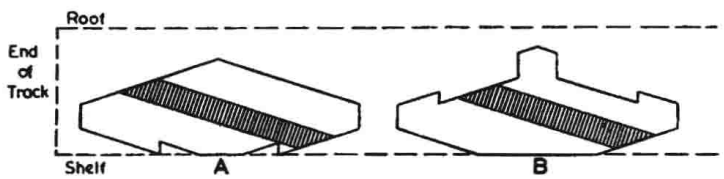


Fig. 1. Elements *A* and *B* in neutral positions on the track

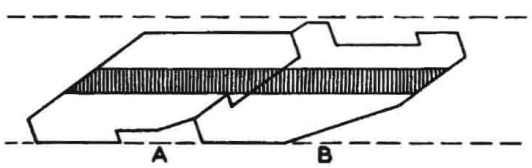


Fig. 2. Elements *A* and *B* hooked together

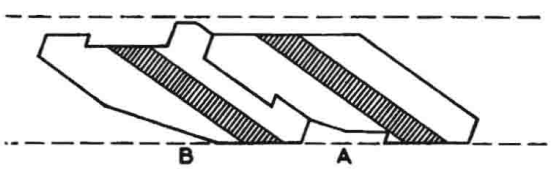
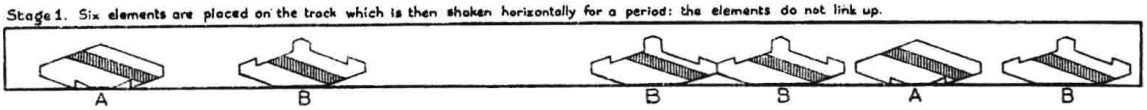


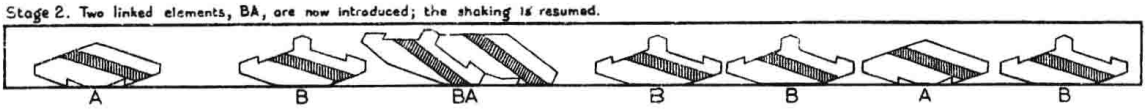
Fig. 3. Alternative complex of *B* and *A*

Model, to show self-replication, made with two types of element, A and B.

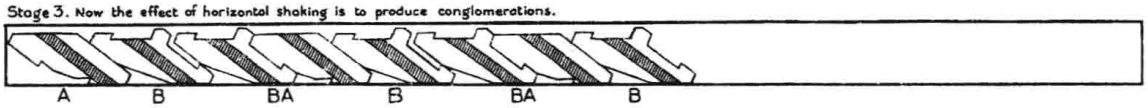
Stage 1. Six elements are placed on the track which is then shaken horizontally for a period: the elements do not link up.



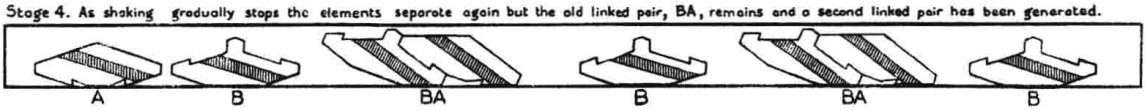
Stage 2. Two linked elements, BA, are now introduced; the shaking is resumed.



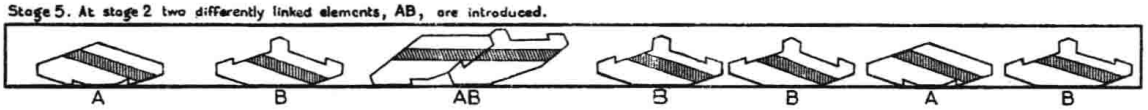
Stage 3. Now the effect of horizontal shaking is to produce conglomerations.



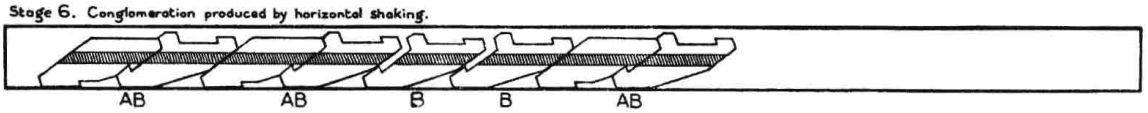
Stage 4. As shaking gradually stops the elements separate again but the old linked pair, BA, remains and a second linked pair has been generated.



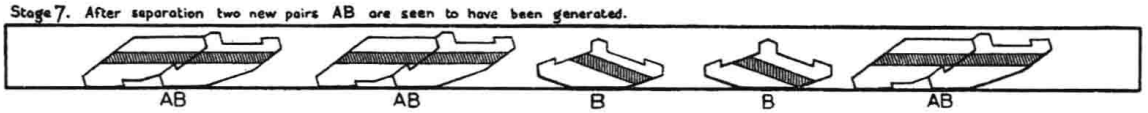
Stage 5. At stage 2 two differently linked elements, AB, are introduced.



Stage 6. Conglomeration produced by horizontal shaking.



Stage 7. After separation two new pairs AB are seen to have been generated.



This diagram supplements the description given by L. S. & R. Penrose in *Nature* **179** (1957) 1183.

9
IMPOSSIBLE OBJECTS: A SPECIAL TYPE OF VISUAL ILLUSION

Brit. J. Psychol. **49** (1958), 31–33. With L. S. Penrose

This article was stimulated by my visit to an exhibition of M. C. Escher's work at the van Gogh Museum in Amsterdam during the International Congress of Mathematicians in 1954. See Chapter 168. After I had drawn the triangle, my father picked up on it and produced, among other things, the impossible staircase shown here. (We were both unaware of numerous other much earlier artistic renderings of such 'impossible objects').

IMPOSSIBLE OBJECTS: A SPECIAL TYPE OF VISUAL ILLUSION

By L. S. PENROSE and R. PENROSE

(*University College, London, and Bedford College, London*)

Two-dimensional drawings can be made to convey the impression of three-dimensional objects. In certain circumstances this fact can be used to induce contradictory perceptual interpretations. Numerous ideas in this field have been exploited by Escher (1954). The present note deals with one special type of figure. Each individual part is acceptable as a representation of an object normally situated in three-dimensional space; and yet, owing to false connexions of the parts, acceptance of the whole figure on this basis leads to the illusory effect of an impossible structure. An elementary example is shown in Fig. 1. Here is a perspective drawing, each part of which is accepted as representing a three-dimensional rectangular structure. The lines in the drawing are, however, connected in such a manner as to produce an impossibility. As the eye pursues the lines of the figure, sudden changes in the interpretation of distance of the object from the observer are necessary. A more complicated structure, not drawn in perspective, is shown in Fig. 2. As this object is examined by following its surfaces, reappraisal has to be made very frequently.

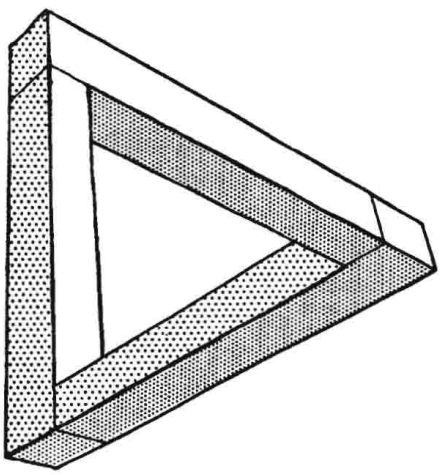


Fig. 1. Perspective drawing of impossible structure.

Another way of presenting the same type of illusion is to express the impossibility in terms of such a phenomenon as a continually descending or ascending path. The flight of steps drawn in Fig. 3 is an example of this. Each part of the structure is acceptable as representing a flight of steps but the connexions are such that the picture, as a whole, is inconsistent; the steps continually descend in a clockwise direction.

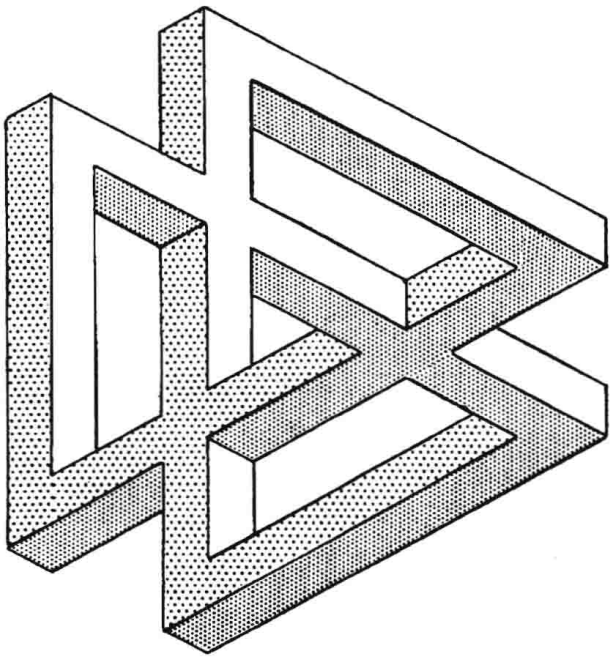


Fig. 2. Diagram of structure with multiple impossibilities.

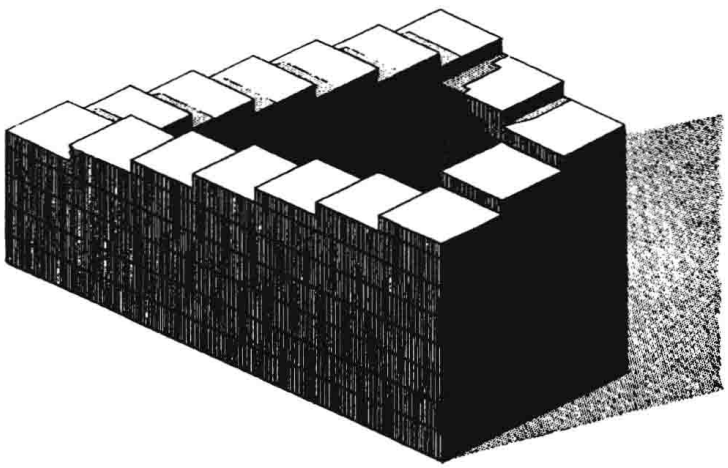


Fig. 3. Continuous flight of steps: shadowed drawing.

Actual objects suitably designed, when viewed from particular angles, can give exactly the same impressions as inconsistent drawings. A photograph of a model of this kind, apparently an impossible staircase, is shown as Fig. 4. Actually the extreme right hand step was much nearer to the camera and on a much higher level than the step which appears to be just above it. Illusions with a different purpose, constructed in a somewhat similar manner, have been discussed by Kilpatrick (1952).

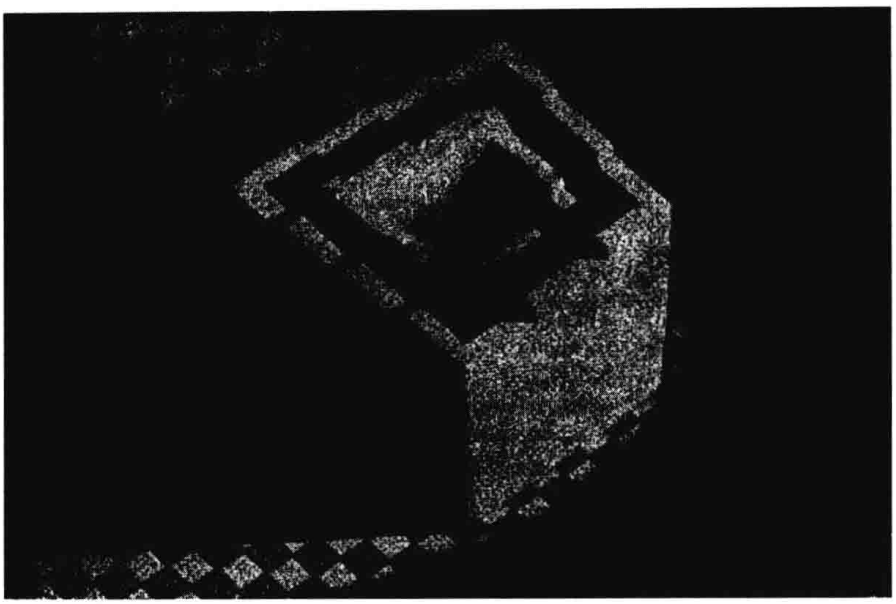


Fig. 4. Untouched photograph of continuous steps: the actual object is rectilinear and the surfaces of the steps are horizontal.

References

ESCHER, M. C. (1954). *Catalogus* 118; Stedelijk Museum, Amsterdam.

KILPATRICK, F. P. (1952). Elementary demonstrations of perceptual phenomena. In *Human Behavior from the Transactional Point of View.* Hanover, N.H.: Institute for Associated Research.

(*Manuscript received* 30 *November* 1956)

10
PUZZLES FOR CHRISTMAS

New Scientist (1958), pp. 1580, 1581, 1597, 1598.
With L. S. Penrose

Puzzles for Christmas

by Professor L. S. PENROSE, F R S, and Dr. ROGER PENROSE

In presenting the reader with a series of problems to provide mental stimulation during the academic vacation, there has been no attempt to provide a continuity of theme. The puzzles are arranged in approximate order of difficulty and the only property they have in common with one another is that they have all been designed by the writers and have not been previously published.

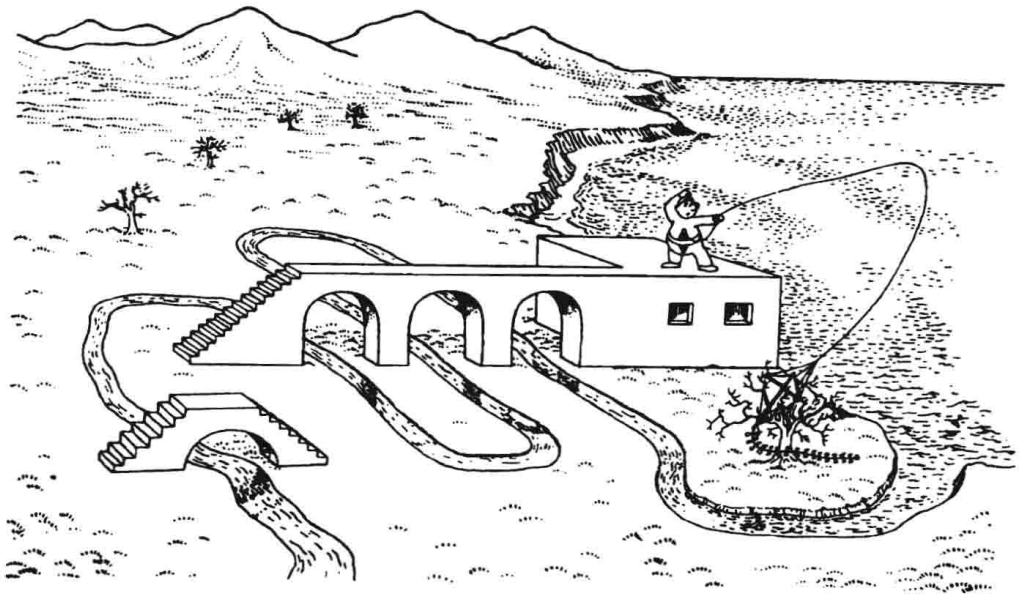


FIGURE 4.

4.—Rescuing a kite

Mr. Tan has been flying a kite in a variable wind and it has come down in a tree, as shown in Figure 4. To retrieve it he must reach the tree on foot carrying his reel of string, of which he has plenty, which is now tucked in his belt. The string must not be tangled, and not caught behind any of the arches, when he reaches the tree. He can only cross the river by the bridge or the viaduct and he can only reach the ground by walking down the steps. Trace the path Mr. Tan must take in order to rescue his kite.

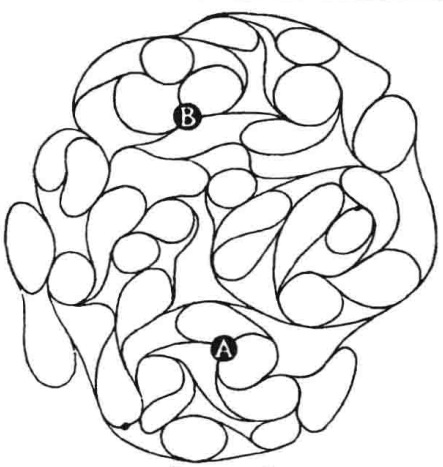


FIGURE 5.

5.—Railway mazes

The following type of maze offers a wide scope for possibilities and it does not appear to have been exploited before. Two examples are shown in Figures 5 and 6. The problem in each case is to get from *A* to *B* by travelling along the lines but without turning any sharp corners and without stopping and reversing Thus, in Figure 7, the routes from x to y or from y to z are allowed but not the one from x to z. The lines may be thought of as representing railway tracks along which an engine with no reverse gear is travelling.

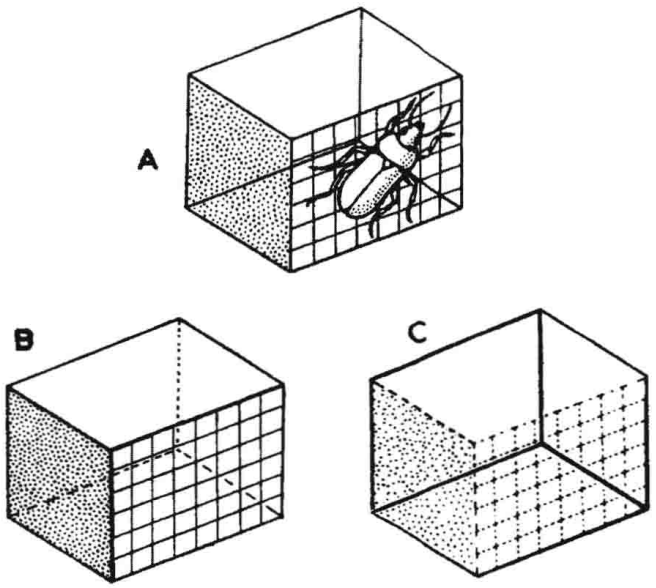


FIGURE 1.

1.—Catching the beetle

This is hardly a puzzle; it is more in the nature of a visual exercise. In Figure 1, a beetle is seen creeping on the vertical side of a rectangular box, *A*, which is similar to *B*. The problem is to get the beetle inside the box, *A*. This is done by covering up *B* and looking at another box, *C*. In a box similar to *C*, the beetle would clearly be sprawling on the floor: but if *C* is covered up, it quickly escapes.

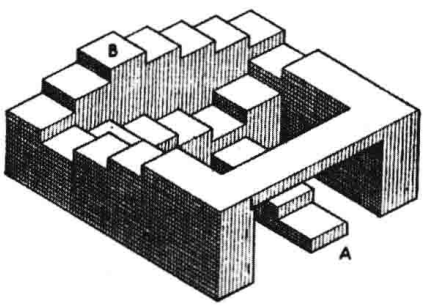


FIGURE 2.

2.—Staircase for lazy people

The first problem in Figure 2 is to go from *A* to *B*, mounting the fewest number of steps. If we turn to the right after four steps and then to the left we climb eleven steps before reaching the top. Similarly, by turning to the left after four steps and to the right after four more, eleven steps are climbed. But *B* can be reached by mounting only ten steps and also descending three on the way. This is done by turning to the right after four steps and then to the right again and then crossing the bridge. The further problem is to find out why this is possible or, rather, why the staircase is impossible.

3.—Parking a car

The principle of impossible objects can be extended to cover situations in which perspective produces peculiar effects because size relationships are rendered inconsistent. In Figure 3, the driver of a motor van wishes to exit through the archway at *A*. At present his van is too large to get through but, by circling the track, never going up hill, in an anticlockwise direction, he can reduce his size sufficiently to get out. The track might be used for making cars smaller so that a great number could be conveniently parked in a confined space. If the reader is not able to follow this argument he should not waste time considering it but should at once proceed to the next puzzle.

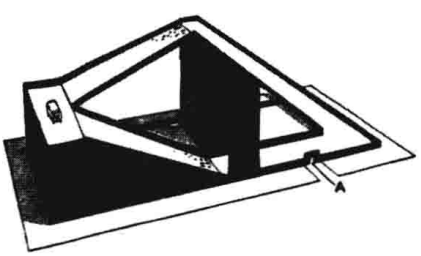


FIGURE 3.

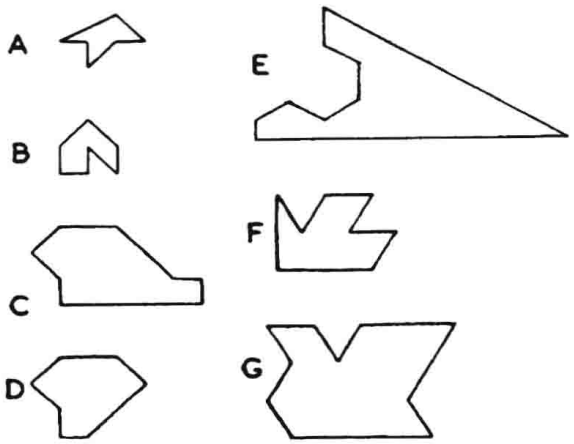


FIGURE 8.

6.—Plane tesselations

It is well known that it is possible to cover a plane with an infinite repeating pattern of squares, or of equilateral triangles or regular hexagons. There are, however, many other shapes which can also be built into such a repeating pattern. For example, for the shape *A* of Figure 8, there is the pattern shown in Figure 9. The reader may care to amuse himself by trying to do the same for the shapes, *B*, *C*, *D*, *E*, *F* and *G*. In Figure 9 the shape *A* is forced to take up four different orientations, two of which are *A* turned over, that is, reflected. How many orientations must be used in the corresponding patterns for *B*, *C*, *D*, *E*, *F* and *G*? In each case the solution is unique.

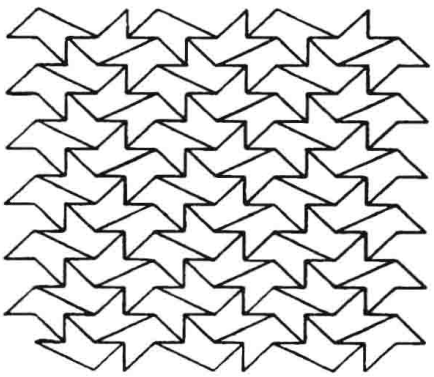


FIGURE 9.

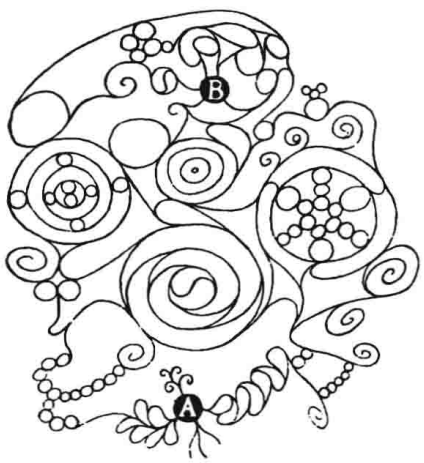


FIGURE 6.

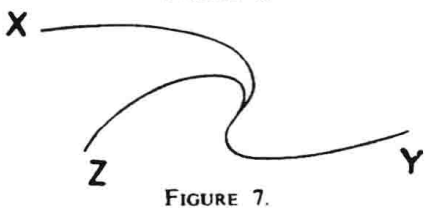


FIGURE 7.

7.—The billiard table problem

This is a problem for the mathematically inclined reader. The puzzle is to design a shape for a billiard table with the property that there are two regions, *A* and *B*, on the table such that it is impossible to hit a ball from any point inside *A* to one inside *B*. The table is to be without pockets or corners, that is to say its shape is a smooth closed curve. It is to be assumed, of course, that the angle made by the ball at the cushion on impact is exactly equal to the corresponding angle made on the rebound, and that there is no friction. Is it possible to extend this result to three dimensions? That is, can we find a smooth closed reflecting surface (mirror) which contains two regions with the property that a source of light placed in one cannot be seen from the other?

SOLUTIONS OVERLEAF

Solutions to puzzles

by L. S. and R. PENROSE

IN puzzle 1 the phenomenon of fluctuation in visual illusions is used for what may be an unexpected purpose. Some people may attain the desired effect by turning the page upside down.

Puzzles 2 and 3 only require the comment that they are constructed on the principle that, by itself, each part of the drawing is logical but that the parts are connected in such a way that the whole figure is inconsistent.

For puzzle 4, the path Mr. Tan must follow is shown by the arrows on the plan in Figure 1 on this page.

The railway mazes in puzzle 5 are solved by tracks as shown in the diagrams on this page: Figure 2 is the solution to the first maze, Figure 3 to the second.

The answers to puzzle 6 concerning tesselations can be tabulated as follows:

Shape	Number of Orientations		
	Original	Reflected	Total
A	2	2	4
B	4	0	4
C	4	4	8
D	4	0	4
E	6	6	12
F	4	4	8
G	6	6	12

The conditions laid down in puzzle 7 for a new shape of billiards table are satisfied by the curved outline shown in Figure 4. The upper arc *PTS* is half of an ellipse whose major axis is *PS* and whose foci are *Q* and *R*. The precise shape of the lower curve is unimportant but it must lie below the major axis and touch it at *Q* and *R*; it must also be smooth and fit smoothly on to the elliptical arc at *P* and *S*.

When a ball leaves the region *A*, it will always bounce back into *A*; thus, it can never reach *B*. The shape can also be modified so as to take into account the finite size of the ball.

If the curve is rotated about the minor axis *TU* of the ellipse, it describes a closed surface in three dimensions which can be shown to have the required property that a source of light placed in the volume traced out by *A* cannot be seen, by any number of reflections, from any point within the volume traced out by *B* (and *vice versa*).

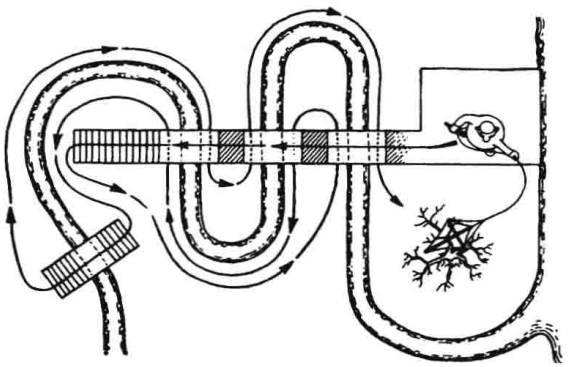
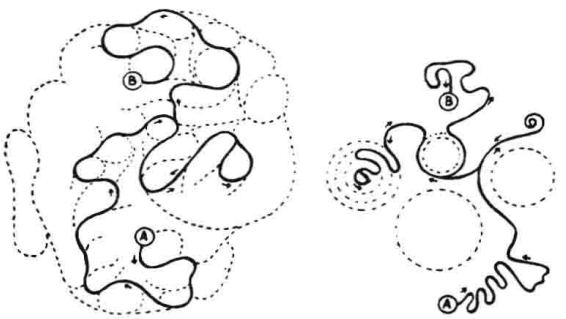


FIGURE 1.



FIGURES 2 *and* 3.

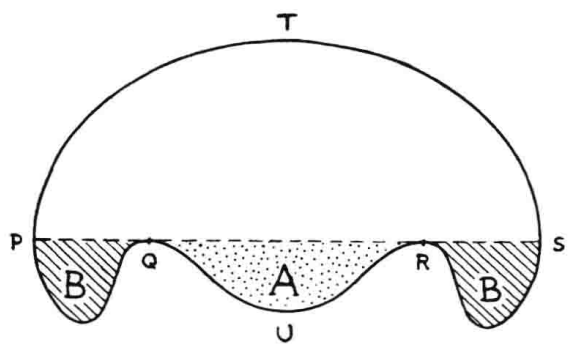


FIGURE 4.

St. Clements Press Ltd., Portugal Street, Kingsway, London, W.C.2.

11
THE APPARENT SHAPE OF A RELATIVISTICALLY MOVING SPHERE

Proc. Camb. Phil. Soc. **55** (1959), 137–139

This was my first published entry into the area of spinors and relativity theory. James Terrell, independently, and using quite different methods, found essentially the same result less than a year later (see Terrell, J. (1959) The invisibility of the Lorentz contraction. *Phys. Rev.* 116, 1041–5).

Reprinted from the *Proceedings of the Cambridge Philosophical Society*,
Volume 55, Part 4, pp. 137–139, 1959.
PRINTED IN GREAT BRITAIN

THE APPARENT SHAPE OF A RELATIVISTICALLY MOVING SPHERE

By R. PENROSE

Received 29 July 1958

It would be natural to assume that, according to the special theory of relativity, an object moving with a speed comparable with that of light should *appear* to be flattened in the direction of motion on account of its FitzGerald–Lorentz contraction. It will be shown here, however, that this is by no means generally the case. It turns out, in particular, that the appearance of a sphere, no matter how it is moving, is always such as to present a *circular* outline to any observer. Thus an instantaneous photograph* of a rapidly moving sphere has the same outline as that of a stationary sphere.

This result may seem paradoxical at first. For example, it might be thought that for a distant sphere moving perpendicularly to the line joining its centre to the observer, the flattening in its direction of motion would certainly be apparent. As the tangents from the observer to the flattened sphere are all nearly the same length, it might seem that the finite velocity of light is irrelevant here. However, the light which appears to the observer to be coming from the leading part of the sphere leaves the sphere at a later time, in the observer's frame, than that which appears to come from the trailing part. (In fact the light from the trailing part reaches the observer from *behind* the sphere, which it can do since the sphere is continually moving out of its way.) The length of the image of the sphere in the direction of motion is thus greater than might otherwise be expected, so that if it were not for the flattening the sphere would appear to be elongated.

In order to prove the exact result that the sphere always presents a circular outline, it is more convenient to consider the sphere as being at rest and the observer moving, this being allowable according to the special principle of relativity. If the sphere is accelerating, in order that it may reasonably be said to remain a sphere throughout its motion, it must be always instantaneously at rest and spherical in *some* Lorentz frame. The sphere will then satisfy the Born conditions for a rigid body (see Synge (3), p. 36). Now the light received by the observer, O, at an instant of his time, from what appears to him to be the outline of the sphere, comes from the sphere at one instant of *its* time. This shows that the acceleration of the sphere is irrelevant. Furthermore, a stationary observer at O clearly sees something with a circular outline. All this is evident from the symmetry. It is therefore only necessary to consider what transformation of the field of vision must be employed when passing from a stationary to a moving observer at the same point, and to show that this transformation is one which sends apparent circles into apparent circles.

* This is not a 'snapshot' in the sense of Synge ((3), p. 120) in which the sphere would indeed appear flattened. I am concerned, here, with a world-picture rather than a world-map (see Milne (1), p. 107), so that the finite velocity of the light coming from the sphere must be taken into account.

One way of doing this is to use the relativistic aberration formula*

$$\tan \tfrac{1}{2}\theta' = \tan \tfrac{1}{2}\theta \sqrt{\frac{c-v}{c+v}}.$$

A stereographic projection of the unit sphere with centre O, from the point for which $\theta = \pi$, sends circles into circles, and the above formula leads simply to an expansion of the plane of projection by an amount $\sqrt{\{(c-v)/(c+v)\}}$.

Alternatively, the following purely geometrical (space-time) argument, suggested to me by J. L. Synge, may be used. Consider the intersection of the past null cone of the event O with any hyperplane through O containing a time-like vector. The hyperplane represents the history of a plane moving with a constant velocity and the past null cone of O represents the history of a sphere with fixed centre and uniformly decreasing radius. They intersect in a circle which converges on O along a right circular cone. Hence, a stationary observer at the event O sees a circle only. Now this holds equally well in any Lorentz frame and therefore shows that a cone of light which appears as a circle to a stationary observer at O also appears as a circle to any moving observer at O.

There is yet another way of obtaining this result, namely by using properties of two-component spinors. Any past null vector (t, x, y, z) with $t < 0, t^2-x^2-y^2-z^2 = 0$ can be represented as a hermitian matrix which is minus the product of a two-component spinor (a, b) with its conjugate (see Veblen (4)). Thus,

$$\begin{pmatrix} t-z & x+iy \\ x-iy & t+z \end{pmatrix} = -\begin{pmatrix} \bar{a}a & \bar{a}b \\ \bar{b}a & \bar{b}b \end{pmatrix}.$$

A direction along the past light cone is then uniquely associated with the corresponding spinor ray, i.e. with the ratio $\lambda = b/a$. Now the points of the light cone at time $t = -1$ constitute a sphere $\mathscr{S}$, whose equations are

$$x^2+y^2+z^2 = 1; \quad t = -1.$$

The field of vision of a stationary observer at the origin may be conveniently represented by this sphere. But $\mathscr{S}$ may be projected stereographically from the point $(-1, 0, 0, -1)$ into the plane $z = 0, t = -1$, with the point $(-1, x, y, z)$ projecting into $(-1, x/(1+z), y/(1+z), 0)$. This plane may be taken as the Argand plane of the complex number

$$\frac{x}{1+z}+i\frac{y}{1+z} = \frac{x+iy}{z-t} = -\frac{\bar{a}b}{\bar{a}a} = -\lambda.$$

The sphere $\mathscr{S}$ can then be regarded as the Argand sphere of $-\lambda$. But a proper Lorentz transformation of the directions through the origin corresponds to a linear transformation of the spinors (a, b), i.e. to a bilinear transformation of λ. This sends circles into circles on $\mathscr{S}$ as required, so that the result follows.

* This is not the usual formula, which is $\cos\theta = \dfrac{\cos\theta' - v/c}{1-(v/c)\cos\theta'}$ (see Synge (3), p. 147). The two forms are easily shown to be equivalent.

It is perhaps worth remarking here that the fact that any proper homogeneous Lorentz transformation is determined by its effect on any three null directions (see Synge (3), p. 99) is here seen to be equivalent to the corresponding property for complex numbers under bilinear transformations (or points of a projective line under a projectivity).

These considerations can also be applied to non-spherical objects moving with uniform velocity. The appearance of such an object is always a circular transform (i.e. product of inversions) of what it would appear in some orientation when stationary. Thus, straight lines appear circular (or straight). Since a stationary circle may appear elliptical, a moving circle can appear boomerang shaped (inverse of an ellipse). In view of this, it is doubtful whether it would be correct to say that a sphere always appears *spherical* rather than just saying it has a circular outline, since the intersection of two spheres does not necessarily appear circular. Nevertheless, there is no other consistent shape which presents a circular outline when stationary, to any observer.

Finally, it may be remarked that all the above considerations apply equally well to a de Sitter space (Minkowski 4-sphere, see Schrödinger (2)) owing to its symmetry. The foregoing arguments apply almost without change.

REFERENCES

(1) Milne, E. A. *Relativity, gravitation and world-structure* (Oxford, 1935).
(2) Schrödinger, E. *Expanding universes* (Cambridge, 1956).
(3) Synge, J. L. *Relativity: the special theory* (Amsterdam, 1956).
(4) Veblen, O. Geometry of two-component spinors. *Proc. Nat. Acad. Sci., Wash.*, 19 (1933), 100–9.

St John's College
Cambridge

12
A SPINOR APPROACH TO GENERAL RELATIVITY

Annals of Physics **10** (1960), 171–201 (reprinted in *General Theory of Relativity* by C. W. Kilmister, Pergamon Press, Oxford, 1973)

The beautiful complete symmetry of the 'gravitational spinor' $\psi_{ABCD'}$ and its attendant properties, was mainly what enticed me irrevocably into the subject of general relativity. But I have regretted confusingly using the letter '∂' for covariant derivative in this particular paper, rather than the more appropriate '∇'. Note the glaring misprint in equation (2.5) and the curious aberration in the displayed expression in the middle of p. [190].

A Spinor Approach to General Relativity

ROGER PENROSE
St. John's College, Cambridge, England

A Spinor Approach to General Relativity

ROGER PENROSE

*St. John's College, Cambridge, England**

A calculus for general relativity is developed in which the basic role of tensors is taken over by spinors. The Riemann–Christoffel tensor is written in a spinor form according to a scheme of Witten. It is shown that the curvature of empty space can be uniquely characterized by a totally symmetric four-index spinor which satisfies a first order equation formally identical with one for a zero rest-mass particle of spin two. However, the derivatives used here are covariant, so that on iteration, instead of the usual wave equation, a nonlinear "source" term appears. The case when a source-free electromagnetic field is present is also considered. (No quantization is attempted here.)

The "gravitational density" tensor of Robinson and Bel is obtained in a natural way as a striking analogy with the spinor expression for the Maxwell stress tensor in the electromagnetic case. It is shown that the curvature tensor determines four gravitational principal null directions associated with flow of "gravitational density", which supplement the two electromagnetic null directions of Synge. The invariants and Petrov type of the curvature tensor are analyzed in terms of these, and a natural classification of curvature tensors is given.

An essentially coordinate-free method is outlined, by which any analytic solution of Einstein's field equations may, in principle, be found. As an elementary example the gravitational and gravitational-electromagnetic plane wave solutions are obtained.

1. INTRODUCTION

An essentially coordinate-free attitude to general relativity will be adopted here. The tensors and spinors occurring are best thought of not as sets of components, but as geometric objects subject to certain formal rules of manipulation. A spinor formalism will be used instead of the usual tensor one, spinors appearing to fit in with general relativity in a remarkably natural way. This adds to a belief that spinors are basically simpler and perhaps more deep-rooted than tensors.

The usual correspondence between tensors and spinors (*1*, *2*) is obtained by the use of a mixed quantity† $\sigma_\mu{}^{AB'}$ satisfying the equation

* Present address: Palmer Physical Laboratory, Princeton University, Princeton, New Jersey.

† For each of the four values of μ, $\sigma_\mu^{AB'}$ is a (2 x 2) Hermitian matrix.

$$\sigma_\mu{}^{A}{}_{C'}\sigma_\nu{}^{BC'} + \sigma_\nu{}^{A}{}_{C'}\sigma_\mu{}^{BC'} = g_{\mu\nu}\epsilon^{AB}, \tag{1.1}$$

where ϵ^{AB}, together with ϵ_{AB}, $\epsilon^{A'B'}$, and $\epsilon_{A'B'}$ is a skew-symmetric "metric" spinor for the 2-dimensional complex spin space. The components of the ϵ's may be taken as $\pm 1, 0$. (To raise or lower a spinor index, one of the ϵ's must be used, e.g., $\xi^A = \epsilon^{AB}\xi_B$, $\xi_B = \xi^A\epsilon_{AB}$.) Primed indices* refer to the complex conjugate spin space. Roman capitals are used here for spinor indices and Greek letters for tensor indices. The spinor equivalent of any tensor is a quantity which has an unprimed and a primed spinor index replacing each tensor index. For example, for a tensor $X^{\lambda\mu}{}_\nu$, we have

$$X^{\lambda\mu}{}_\nu \leftrightarrow X^{AB'CD'}{}_{EF'},$$

where

$$X^{AB'CD'}{}_{EF'} = \sigma_\lambda{}^{AB'}\sigma_\mu{}^{CD'}X^{\lambda\mu}{}_\nu\sigma^\nu{}_{EF'}$$

and

$$X^{\lambda\mu}{}_\nu = \sigma^\lambda{}_{AB'}\sigma^\mu{}_{CD'}X^{AB'CD'}{}_{EF'}\sigma_\nu{}^{EF'},$$

(with $\sigma^\mu{}_{AB'} = g^{\mu\nu}\sigma_\nu{}^{CD'}\epsilon_{CA}\epsilon_{D'B'}$). We have

$$g_{AB'CD'} = \epsilon_{AC}\epsilon_{B'D'}, \quad \delta^{AB'}_{CD'} = \delta^A_C\delta^{B'}_{D'}, \quad g^{AB'CD'} = \epsilon^{AC}\epsilon^{B'D'}. \tag{1.2}$$

The algebraic tensor operations can now all be interpreted as spinor operations. Also the notions of reality of tensors, and of complex conjugate, are interpreted in spinor form with

$$\bar{X}^{\lambda\mu}{}_\nu \leftrightarrow \bar{X}^{A'BC'D}{}_{E'F}$$

so that the roles of primed and unprimed indices are interchanged.[1] Thus reality of tensors is expressed as a Hermitian property of the corresponding spinors.

In addition to the usual correspondence between tensors and spinors given above, there is also a well-known correspondence between real skew-symmetric second rank tensors and symmetric second rank spinors (*2*). Thus if $F_{\mu\nu}$ is real and skew-symmetric, we have

$$F_{AB'CD'} = \tfrac{1}{2}\{\phi_{AC}\epsilon_{B'D'} + \epsilon_{AC}\bar{\phi}_{B'D'}\}, \tag{1.3}$$

where ϕ_{AB} is a uniquely defined symmetric spinor. The right-hand side of (1.3) expresses $F_{AB'CD'}$ as the sum of the part symmetric in A, C (and therefore skew in B', D') and the part skew in A, C (and symmetric in B', D'). (Any skew

* Primed indices are used here rather than the more usual dotted indices, for typographical reasons.

[1] Many authors would omit the bar on the right-hand side. The choice here here is made for reasons of clarity.

pair of spinor indices may be split off as an ϵ-spinor.) A corresponding procedure can be applied to *any* skew-symmetric pair of tensor indices. A tensor with r skew-symmetric pairs of indices thus gives rise to 2^r spinors each with r symmetric pairs of indices in a decomposition similar to (1.3). If the tensor is real, these spinors are paired off as complex conjugates. For an example, see (2.2).

If the tensor $H_{\mu\nu}$ "dual" to $F_{\mu\nu}$ is defined by

$$H_{\mu\nu} = \tfrac{1}{2}\sqrt{-g}F^{\rho\sigma}\epsilon_{\mu\nu\rho\sigma}, \tag{1.4}$$

we have

$$H_{AB'CD'} = \tfrac{1}{2}\{-i\phi_{AC}\epsilon_{B'D'} + i\epsilon_{AC}\phi_{B'C'}\}, \tag{1.5}$$

since if

$$\mathcal{E}_{\mu\nu}^{\alpha\beta} = \sqrt{-g}\epsilon_{\rho\sigma\mu\nu}g^{\alpha\rho}g^{\beta\sigma},$$

then

$$\mathcal{E}_{EF'GH'}^{AB'CD'} = i\delta_E^A\delta_G^C\delta_H^{B'}\delta_{F'}^{D'} - i\delta_G^A\delta_E^C\delta_{F'}^{B'}\delta_{H'}^{D'}. \tag{1.6}$$

(Actually, formulas (1.5) and (1.6) are only correct for one class of choices of $\sigma_\mu{}^{AB}$ satisfying (1.1). If $\sigma_\mu{}^{AB}$ had been chosen from the other class of solutions, the signs of the right-hand sides of (1.5) and (1.6) would be reversed. It will be supposed that the $\sigma_\mu{}^{AB}$ have, in fact, been selected from the appropriate class.) In a similar way any tensor possessing a pair of skew-symmetric indices may be "dualized" with respect to that pair of indices. The spinor decomposition of the "dualized" tensor then differs from that of the original tensor in that the relevant ϵ_{AC} and $\epsilon_{B'D'}$ are, respectively, multiplied by i and by $-i$. This again follows from (1.6). For an example, see (2.6).

General relativity requires, in addition to algebraic properties of tensors, the notion of covariant derivative. The symbol ∂_μ, or correspondingly $\partial_{AB'}$, will be used here to denote *covariant* differentiation. The covariant derivatives of $g_{\mu\nu}$ and of $\sigma_\mu{}^{AB'}$ are both required to be zero.[2] This implies that

$$\partial_\mu\{\epsilon_{AB}\epsilon_{C'D'}\} = 0$$

(see *2*). The stronger conditions

$$\partial_\mu\epsilon_{AB} = 0 \quad \text{and} \quad \partial_\mu\epsilon_{A'B'} = 0 \tag{1.7}$$

will be adopted here (*3*). This enables one to raise and lower spinor indices under the derivative symbol, but it precludes the use of phase transformations of the spinors to generate the electromagnetic field. However, the electromagnetic field

[2] "Spin affinities" $\Gamma^A{}_{B\mu}$, $\bar{\Gamma}^{A'}{}_{B'\mu}$ are introduced to deal with the spinor indices. The conditions (1.7) imply that these spin affinities can be expressed explicitly in terms of $\sigma_\mu{}^{AB'}$ and its coordinate derivatives (see Ruse (*3*)).

will appear here as being associated with spinor transformations in a different way (see 3.13). These two procedures do not appear to combine in an altogether natural way. The simplest formalism, when charges are not present, seems to be obtained when such phase transformations are not permitted.

The point of view adopted here is nearer to that of Rainich (*4*) and of Misner and Wheeler (*5*) in which the electromagnetic field is obtained from the curvature of space-time alone. These phase transformations would not be related in any way to the geometry of the space-time.

2. THE CURVATURE SPINORS

Since the symbol ∂_μ here stands for covariant differentiation, we have

$$\partial_\mu\partial_\nu \not\equiv \partial_\nu\partial_\mu ,$$

the commutation of two ∂'s giving rise to the Riemann–Christoffel tensor $R_{\mu\nu\rho\sigma}$. In fact, we have

$$\{\partial_\mu\partial_\nu - \partial_\nu\partial_\mu\}X_\rho = R_{\nu\mu\rho\sigma}X^\sigma. \tag{2.1}$$

The tensor $R_{\mu\nu\rho\sigma}$ is skew-symmetric in μ, ν and in ρ, σ. Thus, following Witten (*6*), we can apply the procedure outlined in Section 1 and obtain

$$\begin{aligned} R_{AE'BF'CG'DH'} = \tfrac{1}{2}\{&\chi_{ABCD}\epsilon_{E'F'}\epsilon_{G'H'} + \epsilon_{CD}\phi_{ABG'H'}\epsilon_{E'F'} \\ &+ \epsilon_{AB}\bar{\phi}_{E'F'CD}\epsilon_{G'H'} + \epsilon_{AB}\epsilon_{CD}\bar{\chi}_{E'F'G'H'}\}. \end{aligned} \tag{2.2}$$

The spinors χ_{ABCD} and ϕ_{ABGH} are the uniquely defined *curvature spinors*. However, this differs from Witten's form by a factor $\frac{1}{2}$ which is included here for reasons of convenience. From the symmetries of $R_{\mu\nu\rho\sigma}$, it follows that

$$\chi_{ABCD} = \chi_{BACD} = \chi_{ABDC} = \chi_{CDAB} \tag{2.3}$$

and

$$\phi_{ABC'D'} = \phi_{BAC'D'} = \phi_{ABD'C'} = \bar{\phi}_{C'D'AB} . \tag{2.4}$$

Let the right dual $S_{\mu\nu\rho\sigma}$ of $R_{\mu\nu\rho\sigma}$ be defined by

$$S_{\mu\nu\rho\sigma} = \tfrac{1}{2}\sqrt{-g}R_{\mu\nu}{}^{\alpha\beta}\epsilon_{\alpha\beta\rho\sigma} . \tag{2.5}$$

Then from (1.6), we have

$$\begin{aligned} S_{AE'BF'CG'DH'} = \frac{i}{2}\{&-\chi_{ABCD}\epsilon_{E'F'}\epsilon_{G'H'} + \epsilon_{CD}\phi_{ABG'H'}\epsilon_{E'F'} - \epsilon_{AB}\bar{\phi}_{E'F'CD}\epsilon_{G'H'} \\ &+ \epsilon_{AB}\epsilon_{CD}\chi_{E'F'G'H'}\}. \end{aligned} \tag{2.6}$$

Now, the symmetry relation $R_{\mu\nu\rho\sigma} + R_{\mu\rho\sigma\nu} + R_{\mu\sigma\nu\rho} = 0$ is equivalent to

$$S_{\mu\nu\rho}{}^{\nu} = 0,$$

so that multiplying (2.6) by $\epsilon^{BD}\epsilon^{F'H'}$ should give zero (see 1.2). Hence,

$$-\chi_{ABC}{}^{B}\epsilon_{E'G'} - \phi_{ACG'E'} + \bar{\phi}_{E'G'CA} + \epsilon_{AC}\bar{\chi}_{E'F'G'}{}^{F'} = 0$$

The ϕ terms cancel by (2.4), so we have

$$\chi_{ABC}{}^{B} = \lambda\epsilon_{AC}\,, \tag{2.7}$$

where λ is *real* and given by

$$\lambda = \tfrac{1}{2}\chi_{AB}{}^{AB} = \tfrac{1}{2}\bar{\chi}_{E'F'}{}^{E'F'}. \tag{2.8}$$

The reality of λ is, in fact, the only thing new we get out of this identity since (2.7) is implied by (2.3) in any case.

The relations (2.3), (2.4), and (2.8) are the only algebraic relations necessarily satisfied by χ_{ABCD} and $\phi_{ABC'D'}$ for a general Riemannian space, since they imply that an $R_{\mu\nu\rho\sigma}$ given by (2.2) has the required symmetry properties. These relations are all to be found in Witten's paper. However, χ_{ABCD} and $\phi_{ABC'D'}$ also satisfy a differential relation obtained from the Bianchi identity

$$\partial_\tau R_{\mu\nu\rho\sigma} + \partial_\rho R_{\mu\nu\sigma\tau} + \partial_\sigma R_{\mu\nu\tau\rho} = 0.$$

This is equivalent to

$$\partial^\sigma S_{\mu\nu\rho\sigma} = 0,$$

i.e., (by 2.6)

$$\partial^{D}{}_{G'}\chi_{ABCD}\epsilon_{E'F'} - \partial_C{}^{H'}\phi_{ABG'H'}\epsilon_{E'F'} + \epsilon_{AB}\partial^{D}{}_{G'}\bar{\phi}_{E'F'CD} - \epsilon_{AB}\partial_C{}^{H'}\bar{\chi}_{E'F'G'H'} = 0.$$

Separating this into the two equations obtained by, respectively, symmetrizing and skew-symmetrizing with respect to A, B, we get

$$\partial^{D}{}_{G'}\chi_{ABCD} = \partial_C{}^{H'}\phi_{ABG'H'} \tag{2.9}$$

and its complex conjugate. The Bianchi identity is therefore equivalent to (2.9).

There are also relations connecting χ_{ABCD} and $\phi_{ABG'H'}$ with covariant second derivatives of spinors, corresponding to the vector relation (2.1). Let ξ_A be an arbitrary spinor field and define

$$X_{PR'QS'} = \xi_P\xi_Q\epsilon_{R'S'}\,. \tag{2.10}$$

Now (2.1) generalizes to (and in fact implies)

$$\{\partial_\mu\partial_\nu - \partial_\nu\partial_\mu\}X_{\rho\sigma} = R_{\nu\mu\rho\alpha}X^{\alpha}{}_{\sigma} + R_{\nu\mu\sigma\alpha}X_{\rho}{}^{\alpha}. \tag{2.11}$$

But $\partial_\mu\partial_\nu - \partial_\nu\partial_\mu$ is skew-symmetric in μ, ν so that the decomposition (1.3) can be applied:

$$\begin{aligned}\partial_{AC'}\partial_{BD'} - \partial_{BD'}\partial_{AC'} \equiv{}& \tfrac{1}{2}\epsilon_{C'D'}\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}\\ &+ \tfrac{1}{2}\epsilon_{AB}\{\partial_{EC'}\partial^{E}{}_{D'} + \partial_{ED'}\partial^{E}{}_{C'}\}.\end{aligned} \tag{2.12}$$

Thus, (2.11) can be split into two equations each of which must hold separately, one symmetric in A, B (and skew in C′, D′) and the other skew in A, B (and symmetric in C′, D′). Also, any skew pair of indices can be split off as an ϵ-spinor and these may be cancelled throughout the equation. Hence by (2.2), the equation symmetric in A, B is

$$\begin{aligned}\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}\xi_P\xi_Q\epsilon_{R'S'} &= \chi_{ABPC}\xi^C\xi_Q\epsilon_{R'S'} \\ &+ \chi_{ABQC}\xi_P\xi^C\epsilon_{R'S'} - \phi_{ABR'C'}\xi_P\xi_Q\delta_{S'}^{C'} + \phi_{ABS'C'}\xi_P\xi_Q\delta_{R'}^{C'}.\end{aligned} \tag{2.13}$$

The ϕ terms cancel and, because of (1.7), the $\epsilon_{R'S'}$ term may be divided out. Also,

$$\partial_\mu\partial_\nu(\xi_P\xi_Q) = \xi_P\partial_\mu\partial_\nu\xi_Q + \xi_Q\partial_\mu\partial_\nu\xi_P + (\partial_\nu\xi_P)(\partial_\mu\xi_Q) + (\partial_\mu\xi_Q)(\partial_\nu\xi_P),$$

whence

$$\{\partial_\mu\partial_\nu - \partial_\nu\partial_\mu\}(\xi_P\xi_Q) = \xi_P\{\partial_\mu\partial_\nu - \partial_\nu\partial_\mu\}\xi_Q + \xi_Q\{\partial_\mu\partial_\nu - \partial_\nu\partial_\mu\}\xi_P\,.$$

It follows that

$$\begin{aligned}\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}(\xi_P\xi_Q) &= \xi_P\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}\xi_Q \\ &+ \xi_Q\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}\xi_P \\ &= \xi_P\chi_{ABQC}\xi^C + \xi_Q\chi_{ABPC}\xi^C\end{aligned}$$

by (2.13). Multiplying this equation by $\eta^P\eta^Q$ where η^A is chosen arbitrarily, we get

$$2(\eta^P\xi_P)(\eta^Q\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}\xi_Q) = 2(\eta^P\xi_P)(\eta^Q\chi_{ABQC}\xi^C).$$

Since η^A is arbitrary, we may divide by $2(\eta^P\xi_P)$ and obtain

$$\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}\xi_Q = \chi_{ABQC}\xi^C. \tag{2.14}$$

Also the equation obtained from (2.11) which is skew in A, B and symmetric in C′, D′ gives rise to

$$\{\partial_{EC'}\partial^E{}_{D'} + \partial_{ED'}\partial^E{}_{C'}\}\xi_Q = \phi_{QAC'D'}\xi^A \tag{2.15}$$

in an exactly similar way. The corresponding results for a primed spinor $\zeta_{A'}$ are obtained by taking the complex conjugates of (2.14) and (2.15). Thus,

$$\{\partial_{EC'}\partial^E{}_{D'} + \partial_{ED'}\partial^E{}_{C'}\}\zeta_{A'} = \bar{\chi}_{C'D'A'B'}\zeta^{B'} \tag{2.16}$$

and

$$\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}\zeta_{C'} = \phi_{ABC'D'}\zeta^{D'}. \tag{2.17}$$

The corresponding relations for spinors with more than one index can be ob-

tained from (2.14), . . ., (2.17) since any spinor can be expressed as a linear combination of products of one-index spinors. Spinors with upper indices present no extra problem because the derivative of an ϵ-spinor is zero. As an example, we have

$$\{\partial_{AF'}\partial_B{}^{F'} + \partial_{BF'}\partial_A{}^{F'}\}\beta_C{}^{DE'} = \chi_{ABCP}\beta^{PDE'} + \chi_{AB}{}^{D}{}_{P}\beta_C{}^{PE'} + \phi_{AB}{}^{E'}{}_{Q'}\beta_C{}^{DQ'}.$$

In particular, by applying this to a "vector" $X^{DE'}$, and using (2.12) and (2.2), we can get back to (2.1). (It is not so easy to obtain (2.14), . . ., (2.17) directly from (2.1) rather than from (2.11), since the fact that the ϵ-spinors are constant must be used somewhere in the argument.)

The geometry of a Riemannian space (with signature $+---$) can thus be described entirely in spinor terms, with the role of the curvature tensor being taken over by spinors χ_{ABCD}, $\phi_{ABE'F'}$ satisfying (2.3), (2.4), (2.8), (2.9), (2.14), (2.15), (2.16), and (2.17).

3. THE EINSTEIN CONDITIONS

The theory of Section 2 will now be specialized to two cases of particular note, namely empty space-time and source-free electromagnetic field.

The Ricci tensor $R_{\mu\nu} = R^{\sigma}{}_{\mu\sigma\nu}$ has the spinor form

$$\begin{aligned} R_{AC'BD'} &= \tfrac{1}{2}\{\chi_{EA}{}^{E}{}_{B}\epsilon_{C'D'} - 2\phi_{ABC'D'} + \epsilon_{AB}\bar{\chi}_{F'C'}{}^{F'}{}_{D'}\} \\ &= \lambda\epsilon_{AB}\epsilon_{C'D'} - \phi_{ABC'D'} \end{aligned}$$

by (2.2), (2.7), (2.8). The scalar curvature $R = R^{\sigma}{}_{\sigma}$ is given by

$$R = 4\lambda \tag{3.1}$$

because of the symmetry of $\phi_{ABC'D'}$. The Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R$ takes the form

$$G_{AC'BD'} = -\lambda\epsilon_{AB}\epsilon_{C'D'} - \phi_{ABC'D'}. \tag{3.2}$$

Einstein's equations $G_{\mu\nu} = 0$ for empty space clearly give

$$\phi_{ABC'D'} = 0 \tag{3.3}$$

and

$$\lambda = 0.$$

On the other hand, if it is required to include a cosmological term in Einstein's equations, we have only $\phi_{ABC'D'} = 0$, the cosmological constant being equal to λ by (3.1).

Supposing for the moment that the cosmological constant is zero, (2.7) gives

$$\chi_{ABC}{}^{B} = 0,$$

that is, χ_{ABCD} is symmetric in B and D. But by (2.3), it is also symmetric A, B and in C, D. It is therefore *completely symmetric* in all its indices.

It is a remarkable and perhaps significant fact, that only for a manifold with the apparently arbitrary $+---$ signature of our space-time, and which satisfies the Einstein equations for empty space, can its curvature be characterized by so natural an object as a totally symmetric four-index spinor. The geometry of this spinor will be dealt with in Section 3.

If a cosmological term (or matter) is present, we can write

$$\chi_{ABCD} = \psi_{ABCD} + \frac{\lambda}{3}\{\epsilon_{AC}\epsilon_{BD} + \epsilon_{AD}\epsilon_{BC}\} \tag{3.4}$$

and then ψ_{ABCD} will be totally symmetric even if $\lambda \neq 0$. The spinor ψ_{ABCD} defined by (3.4) will be called here the *gravitational* spinor (even in cases where $\phi_{ABC'D'} \neq 0$). It corresponds uniquely to Weyl's conformal tensor $C_{\mu\nu\rho\sigma}$.

Relation (2.9) gives (with $\phi_{ABC'D'} = 0$)

$$\partial^{DE'}\psi_{ABCD} = 0 \tag{3.5}$$

and of course $\partial^{DE'}\lambda = 0$ also. Equation (3.5) has the suggestive appearance of being formally identical with a spinor equation for a zero rest-mass particle of spin two. (See Dirac (7) and compare (3.10).) However, the differentiation used here is covariant, so that derivatives do not commute. Hence, new features arise with second and higher derivatives. In particular, it is not true that Eq. (3.5) leads to the covariant wave equation upon iteration with $\partial_{FE'}$. We have

$$\partial_{FE'}\partial_D{}^{E'} \equiv \tfrac{1}{2}\{\partial_{FE'}\partial_D{}^{E'} + \partial_{DE'}\partial_F{}^{E'}\} + \tfrac{1}{2}\epsilon_{FD}\Box, \tag{3.6}$$

where

$$\Box \equiv \partial_\mu\partial^\mu \equiv \partial_{FE'}\partial^{FE'}.$$

Also,

$$\{\partial_{FE'}\partial_D{}^{E'} + \partial_{DE'}\partial_F{}^{E'}\}\xi_A = \psi_{FDAB}\xi^B - \frac{\lambda}{3}\{\xi_D\epsilon_{FA} + \xi_F\epsilon_{DA}\} \tag{3.7}$$

by (2.14) and (3.4). Now (3.6) gives

$$0 = \partial_{FE'}\partial_D{}^{E'}\psi_{ABC}{}^D = \tfrac{1}{2}\{\partial_{FE'}\partial_D{}^{E'} + \partial_{DE'}\partial_F{}^{E'}\}\psi_{ABC}{}^D - \tfrac{1}{2}\Box\psi_{ABCF}.$$

By (3.7), this leads to

$$\begin{aligned}\Box\psi_{ABCD} &= \psi_{ABEF}\psi_{CD}{}^{EF} + \psi_{ACEF}\psi_{DB}{}^{EF} + \psi_{ADEF}\psi_{BC}{}^{EF} - 2\lambda\psi_{ABCD} \\ &= 3\psi_{(AB}{}^{EF}\psi_{CD)EF} - 2\lambda\psi_{ABCD}\end{aligned} \tag{3.8}$$

where the indices between the brackets are to be symmetrized.[3] Thus, even when $\lambda = 0$ there is the nonlinear term on the right. This shows that the ψ-field can perhaps be thought of as acting as its own source to a certain extent. If ψ_{ABCD} is small we have

$$\Box\psi_{ABCD} \doteq 0$$

since λ is small in any case. Equation (3.8) indicates that we can only expect to have *exact* solutions for plane gravitational waves moving with the velocity of light when $\lambda = 0$ and $\psi_{(AB}{}^{EF}\psi_{CD)EF} = 0$. This question will be returned to in Section 4 where this condition will be interpreted geometrically and in Section 5 where such an exact solution will be given.

The tensor $T_{\mu\nu\rho\sigma}$ whose spinor equivalent is given by

$$T_{AE'BF'CG'DH'} = \psi_{ABCD}\bar{\psi}_{E'F'G'H'} \tag{3.9}$$

is of considerable interest. It has the properties of complete symmetry in its tensor indices, vanishing traces (as easily follows from (3.9)) and vanishing covariant divergence (with or without λ-term in Einstein's equations), since by (3.5)

$$\partial^{DH'}\{\psi_{ABCD}\bar{\psi}_{E'F'G'H'}\} = 0.$$

It would therefore appear that $T_{\mu\nu\rho\sigma}$ is a multiple of the "gravitational density" (or "super-energy") tensor due independently to I. Robinson (unpublished seminars) and to Bel (*8*, *9*).[4] As is easily verified, $T_{\mu\nu\rho\sigma}$ is, in fact, proportional to the Robinson-Bel tensor. Equation (3.9) bears a striking resemblance to the corresponding Eq. (3.11) for the electromagnetic case.

Let us now suppose that there is a source-free electromagnetic field present. The field tensor $F_{\mu\nu}$ can be expressed according to (1.3) in terms of a symmetric spinor ϕ_{AB}. This spinor can be used instead of $F_{\mu\nu}$ to represent the electromagnetic field (*2*), and the Maxwell field equations (in covariant form) become

$$\partial^{AC'}\phi_{AB} = 0. \tag{3.10}$$

[3] The tensor form of this relation is

$$\Box R_{\mu\nu\rho\sigma} = R_{\mu\nu}{}^{\alpha\beta}R_{\alpha\beta\rho\sigma} + 4R^{\alpha}{}_{\mu\beta[\rho}R^{\beta}{}_{\sigma]\alpha\nu} - 2\lambda R_{\mu\nu\rho\sigma}\,.$$

[4] This tensor was also found by R. Sachs (*10*) working with the group at Hamburg, and by A. Komar. However, only Robinson noticed the *total* symmetry of the tensor expression. It is not hard to see in the spinor formalism that, if $R_{\mu\nu} = 0$, *any* four-index tensor, quadratic and homogeneous in $R^{\mu}{}_{\nu\rho\sigma}$ and with vanishing divergence, must be totally symmetric. Robinson's tensor expression (with ($S_{\mu\nu\rho\sigma}$ as in (2.5)) is

$$T_{\mu\nu\rho\sigma} = R_{\mu\alpha\beta}{}^{\gamma}R_{\rho}{}^{\alpha}{}_{\sigma}{}^{\beta} + S_{\mu\alpha\nu\beta}S_{\rho}{}^{\alpha}{}_{\sigma}{}^{\beta}.$$

The energy-momentum tensor $T_{\mu\nu}$ for the electromagnetic field is given by

$$T_{AC'BD'} = \tfrac{1}{2}\phi_{AB}\bar{\phi}_{C'D'} \tag{3.11}$$

(see *10*). Now Einstein's equations with cosmological term are

$$G_{\mu\nu} + \lambda g_{\mu\nu} = -\kappa T_{\mu\nu} .$$

The λ defined by (2.8) is still the cosmological constant, because $T^{\nu}{}_{\nu} = 0$ and by (3.1), $4\lambda = R = -G^{\nu}{}_{\nu}$. Choosing units suitably so that $\kappa = 2$ (or absorbing the constant into the definition of ϕ_{AB}) we have, from (3.2), (3.11)

$$\phi_{ABC'D'} = \phi_{AB}\bar{\phi}_{C'D'} .$$

Equation (2.9) now gives

$$\partial^{D}{}_{G'}\psi_{ABCD} = \bar{\phi}_{G'H'}\partial_{C}{}^{H'}\phi_{AB} \tag{3.12}$$

by (3.4) and (3.10) since λ is necessarily constant. Thus the ϕ-field appears as a kind of source term to the ψ-field (here in the first-order equation).

From (2.17) we have

$$\{\partial_{FE'}\partial_{D}{}^{E'} + \partial_{DE'}\partial_{F}{}^{E'}\}\zeta_{G'} = \phi_{DF}\bar{\phi}_{G'H'}\zeta^{H'} \tag{3.13}$$

and (3.7) still holds. The second order equation arising from (3.10) turns out to be

$$\Box\phi_{AB} = \psi_{ABCD}\phi^{CD} - \tfrac{4}{3}\lambda\phi_{AB}$$

so that even the Maxwell field does not exactly satisfy the covariant wave equation (compare Eddington, *11*, Section 74). Also, (3.5) leads to

$$\Box\psi_{ABCD} = 3\psi_{(AB}{}^{EF}\psi_{CD)EF} - 2\lambda\psi_{ABCD} - 2\bar{\phi}_{G'H'}\partial_{A}{}^{G'}\partial_{B}{}^{H'}\phi_{CD}$$

4. THE GEOMETRY AND INVARIANTS OF ψ_{ABCD}

It is known that a general electromagnetic field determines two real principal null directions at each point (*12*). These are given in the general case by the real eigenvectors of the field tensor $F^{\mu}{}_{\nu}$ considered as a matrix. (There are also two complex null directions given by the complex eigenvectors, but these add nothing to the geometry as they are determined by their orthogonality with the real ones.) An alternative method of obtaining these principal null directions is to use a spinor approach. Any null vector x^{μ} corresponds to the product of a dotted with an undotted spinor

$$x^{AB'} = \eta^{A}\theta^{B'} .$$

If x^{μ} is real, $\theta^{B'}$ is a multiple of $\bar{\eta}^{B'}$, positive if x^{μ} points to the future. Any direction along the light cone therefore corresponds uniquely to a one-index spinor

ray (set of spinors proportional to a given spinor). Now $F_{\mu\nu}$ corresponds uniquely to ϕ_{AB} (by 1.3) and we have

$$F^{AC'}{}_{BD'} = -\tfrac{1}{2}\{\phi^{A}{}_{B}\delta^{C'}_{D'} + \delta^{A}_{B}\bar{\phi}^{C'}{}_{D'}\}.$$

It is easily verified from this that the eigenvectors of $F^{AC'}{}_{BD'}$ are $\eta^{A}\bar{\eta}^{B'}$, $\zeta^{A}\bar{\zeta}^{B'}$ (corresponding to the real null vectors) and $\eta^{A}\bar{\zeta}^{B'}$, $\zeta^{A}\bar{\eta}^{B'}$ (corresponding to the complex null vectors) where

$$\phi_{AB} = \tfrac{1}{2}\{\eta_{A}\zeta_{B} + \eta_{B}\zeta_{A}\} = \eta_{(A}\zeta_{B)}\,. \tag{4.1}$$

See also Witten (*13*). A decomposition exactly analogous to (4.1) exists for the gravitational spinor. We have

$$\psi_{ABCD} = \alpha_{(A}\beta_{B}\gamma_{C}\delta_{D)}\,, \tag{4.2}$$

which expresses the gravitational spinor uniquely (except for scale factors) as a symmetrized product of one-index spinors. The bracket here denotes symmetrization as before, so that written out in full, there would be 24 terms on the right-hand side. The existence and uniqueness of (4.2) follows from the fundamental theorem of algebra:

$$\psi_{ABCD}\xi^{A}\xi^{B}\xi^{C}\xi^{D} = (\alpha_{A}\xi^{A})(\beta_{B}\xi^{B})(\gamma_{C}\xi^{C})(\delta_{D}\xi^{D}) \tag{4.3}$$

expresses the general binary quartic form as a product of linear factors. These factors are essentially unique, and equating coefficients gives (4.2).

Now the spinors α_{A}, β_{B}, γ_{C}, δ_{D} determine four directions along the light cone. These are uniquely determined by ψ_{ABCD} and will be called the *gravitational principal null directions*.[5] They supplement the two electromagnetic principal null directions corresponding to η_{A} and ζ_{A}. The gravitational principal null directions are only undefined if $\psi_{ABCD} = 0$ but they may coincide in special cases. In particular, for the case of the Schwarzschild solution, it follows from the symmetry that they must coincide in pairs at every point, one pair pointing towards the origin along the light cone and the other pair pointing away from it. (Time reversal symmetry shows that they cannot all four coincide or coincide three and one.) The coincidence of the two electromagnetic null directions is the condition for the electromagnetic field to be null. (The electromagnetic directions are, of course, only undetermined if $\phi_{AB} = 0$.) Thus, for an electromagnetic plane wave, the principal null directions coincide and, naturally enough, point in the direction of motion of the wave. Similarly, it turns out that for a gravitational plane wave, all the gravitational null directions coincide. This

[5] These four null directions are implicit in the work of Ruse (*14*). They correspond to the self-conjugate lines of the Riemannian complex. *Note added in proof:* They have been further exploited by Debever (*15*, *16*).

question will be returned to later. Gravitational radiation is sometimes analysed in terms of the invariants of the Riemann tensor (*17*) and it will be useful first to relate these invariants to the null directions defined above.

The number of independent invariants of the Riemann tensor in empty space is well known to be four. These may be interpreted as the real and imaginary parts of two independent complex invariants of ψ_{ABCD}, e.g.,

$$I = \psi_{ABCD}\psi^{ABCD}, \quad J = \psi^{AB}{}_{CD}\psi^{CD}{}_{EF}\psi^{EF}{}_{AB} \tag{4.4}$$

(see Witten, *6*, p. 359). These may be thought of as invariants of the binary quartic form (4.3). According to the theory of invariants of binary forms, I and J are independent and any invariant of the quartic form (4.3) is a function of them (see Grace and Young, *18*). Thus the real and imaginary parts of I and J are a complete set of curvature invariants for empty space. The invariants I and J take the following tensor form if $R_{\mu\nu} = \lambda g_{\mu\nu}$:

$$I = \frac{1}{2} R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} + \frac{i}{4}\sqrt{-g}R_{\mu\nu}{}^{\alpha\beta}\epsilon_{\alpha\beta\rho\sigma}R^{\mu\nu\rho\sigma} - \frac{4}{3}\lambda^2,$$

$$J = \frac{1}{2}\left\{R^{\mu\nu}{}_{\rho\sigma} + \frac{i}{2}\sqrt{-g}R^{\mu\nu\alpha\beta}\epsilon_{\alpha\beta\rho\sigma}\right\} R^{\rho\sigma}{}_{\gamma\delta}R^{\gamma\delta}{}_{\mu\nu} - 2\lambda I - \frac{8}{9}\lambda^3$$

with $\lambda = \frac{1}{4}R$. These relations are obtained from (2.2), (2.5), (2.6), and (3.4). For a general curvature tensor,[6] the tensor $R_{\mu\nu\rho\sigma}$ in the above expressions must be replaced by

$$\tfrac{1}{2}R_{\mu\nu\rho\sigma} + \tfrac{1}{8}gR^{\alpha\beta\gamma\delta}\epsilon_{\alpha\beta\mu\nu}\epsilon_{\gamma\delta\rho\sigma}\,.$$

Binary forms have a geometrical interpretation as sets of points on a complex projective line. The equation

$$\psi_{ABCD}\xi^A\xi^B\xi^C\xi^D = 0$$

is satisfied if and only if at least one of the factors $\alpha_A\xi^A$, $\beta_B\xi^B$, $\gamma_C\xi^C$, $\delta_D\xi^D$ vanishes, each of the conditions $\alpha_A\xi^A = 0, \cdots, \delta_D\xi^D = 0$ representing a point on the line. Thus ψ_{ABCD} corresponds to four points A, B, C, D on a complex projective line the coordinates of these points being the components of α_A, β_A, γ_A, δ_A, respec-

[6] It is perhaps worth remarking that a general method of converting expressions involving ψ_{ABCD} into the corresponding expressions for $R_{\mu\nu\rho\sigma}$ would be to use the formula

$$\psi_{ABCD} = \tfrac{1}{2}R_{AE'BF'CG'DH'}\epsilon^{E'F'}\epsilon^{G'H'} - \tfrac{1}{12}R\{\epsilon_{AC}\epsilon_{BD} + \epsilon_{AD}\epsilon_{BC}\}$$

but the conversion of spinor contractions to an equivalent tensor form is sometimes complicated.

tively. Now any four collinear points have a projective invariant, namely, their cross-ratio

$$\mu = \frac{(\alpha_A \beta^B)(\gamma_B \delta^B)}{(\alpha_C \delta^C)(\gamma_D \beta^D)}.$$

This cross-ratio is the only independent invariant of the four points and is therefore the only independent invariant of ψ_{ABCD} which is unchanged if ψ_{ABCD} is multiplied by a non-zero complex number. Thus, the four real invariants of the curvature of empty space can be interpreted as a complex cross-ratio,[7] and a phase and a magnitude[8] for ψ_{ABCD}.

This phase is associated with duality rotations of the curvature tensor (suggested to me first by I. Robinson) which are exactly analogous to electromagnetic duality rotations (*5*). In each case the duality rotation invariance of the first-order equation (3.5), (3.10) is broken only when sources are present. Letting

$$\psi_{ABCD} \rightarrow e^{i\theta} \psi_{ABCD},$$

where θ is a real constant, we have, assuming for simplicity that $\phi_{ABC'D'}$ and λ both vanish,

$$R_{\mu\nu\rho\sigma} \rightarrow \cos\theta\, R_{\mu\nu\rho\sigma} - \sin\theta\, S_{\mu\nu\rho\sigma}$$

by (2.2) and (2.6), $S_{\mu\nu\rho\sigma}$ being the right (or equivalently the left) dual of $R_{\mu\nu\rho\sigma}$ defined by (2.5). This is exactly analogous to

$$\phi_{AB} \rightarrow e^{i\theta} \phi_{AB}$$

giving

$$F_{\mu\nu} \rightarrow \cos\theta\, F_{\mu\nu} - \sin\theta\, H_{\mu\nu}$$

where the dual $H_{\mu\nu}$ of $F_{\mu\nu}$ is given by (1.4). Unlike the electromagnetic case, however, duality rotations of the ψ-field of an empty space solution do not in general give rise to new exact solutions of the field equations. (See, for example, Eq. (3.8).)

It will be observed that the Robinson–Bel tensor $\psi_{ABCD}\bar{\psi}_{E'F'G'H'}$ determines ψ_{ABCD} up to a duality rotation in the same way that $\phi_{AB}\bar{\phi}_{C'D'}$ determines ϕ_{AB}

[7] The idea of using a complex cross-ratio as an invariant defined by four null rays has also been independently suggested by I. Robinson (unpublished).

[8] This phase and magnitude of ψ_{ABCD} can be interpreted in an invariant way as the argument and modulus of, say, $\sqrt{I}$. This is not really satisfactory, however, since I may vanish. It might be better to use the argument and modulus of the κ which is defined by the relations (4.5). This only need vanish if $I = J = 0$, the condition for three of the null directions to coincide. Its definition depends on an arbitrary ordering of the null directions, however, as does the definition of μ.

up to a duality rotation. The principal null directions are therefore associated even more closely with these "energy" expressions than with the field quantities themselves. These expressions are completely characterized by the principal null directions, apart from their actual magnitude. It might be expected that the gravitational null directions are in some way associated with flow of "gravitational density." There does appear to be such a connection, as may be seen from the following argument.

Let x_μ be any null vector pointing into the future, so that

$$x_{AB'} = \xi_A \bar{\xi}_{B'} .$$

Then by (3.9) and (4.3)

$$\begin{aligned} T_{\mu\nu\rho\sigma} x^\mu x^\nu x^\rho x^\sigma &= (\psi_{ABCD}\xi^A\xi^B\xi^C\xi^D)(\bar{\psi}_{E'F'G'H'}\bar{\xi}^{E'}\bar{\xi}^{F'}\bar{\xi}^{G'}\bar{\xi}^{H'}) \\ &= (a_\mu x^\mu)(b_\nu x^\nu)(c_\rho x^\rho)(d_\sigma x^\sigma), \end{aligned}$$

where

$$a_{AB'} = \alpha_A \bar{\alpha}_{B'} , \; b_{AB'} = \beta_A \bar{\beta}_{B'} , \; c_{AB'} = \gamma_A \bar{\gamma}_{B'} , \; d_{AB'} = \delta_A \bar{\delta}_{B'} .$$

The vectors a_μ , b_μ , c_μ , d_μ are null vectors, pointing into the future, corresponding to the gravitational principal null directions. Thus $T_{\mu\nu\rho\sigma} x^\mu x^\nu x^\rho x^\sigma$ only vanishes for null vectors x^μ which point in one of the gravitational principal null directions. Otherwise it is positive. But for any time-like vector t^μ, the expression

$$\frac{T_{\mu\nu\rho\sigma} t^\mu t^\nu t^\rho t^\sigma}{(t_\alpha t^\alpha)^2} \tag{4.5}$$

measures the gravitational density for an observer whose time axis is t^μ (see Bel (*8*, *9*)). It is positive (for empty space) unless $R_{\mu\nu\rho\sigma} = 0$. Thus the gravitational principal null directions are characterized by the fact that for observers travelling with a given velocity infinitessimally less than c, the gravitational density will be a minimum for those observers who travel approximately along a principal null direction.

It is convenient, from a geometrical point of view, to represent null directions as points on a sphere. This sphere may be thought of as the field of vision of some observer. It may also be interpreted as a realization of the complex projective line mentioned above. (A complex projective line is, topologically, a real 2-sphere.) This sphere is the Argand sphere of the ratio of the two components of a one-index spinor (see Penrose, *19*, p. 138). Any Lorentz transformation corresponds to a bilinear transformation of this ratio and therefore to a projective (conformal) transformation of the sphere, which sends circles into circles.

Four points on the sphere are concyclic if and only if their cross-ratio is real. A particular case of this is harmonic points for which the cross-ratio is -1, 2, or

½ according to the order in which the points are taken. The symmetry of a harmonic set is best exhibited when the points are equally spaced around a great circle. The symmetries are then just the symmetries of a square. Any harmonic set can be brought into this form by a suitable projective (Lorentz) transformation, since any three points on the sphere can be transformed into any three others. Harmonic sets are of interest here because they have a greater symmetry than a general set of four points. They correspond to the vanishing of the invariant J (see Grace and Young, *18*, p. 206). Also of interest is the equianharmonic set which has an even greater symmetry. The cross-ratio here is $-\omega$ or $-\omega^2$ where $\omega = e^{i2\pi/3}$. By a suitable projective transformation these four points can be made the vertices of a regular tetrahedron. Equianharmonic points correspond to the vanishing of the invariant I (*18*, p. 206).

In the case of a general cross-ratio μ the symmetry is given by the Klein 4-group (except that there are also some reflectional symmetries if μ is real or has modulus unity). There is a unique projective transformation (involution) which interchanges any pair of the points with the remaining pair. These and the identity constitute the complete projective symmetry group provided that μ is different from $-1, 2, \frac{1}{2}, -\omega, -\omega^2, 0, 1$, or ∞, the cases 0, 1, and ∞ occurring when a pair of points coincide. The value of μ can be obtained from I^3/J^2 since it can be shown that

$$I = 6\kappa^2(\mu + \omega)(\mu + \omega^2), \quad J = 6\kappa^3(\mu + 1)(\mu - \tfrac{1}{2})(\mu - 2) \quad (4.6)$$

for some κ (see (*16*) p. 205). There are in general six values of μ for a given value of I^3/J^2. They correspond to different orders in which the four points can be taken. The values are μ, $1 - \mu$, $1/\mu$, $1 - (1/\mu)$, $1/(1 - \mu)$, $\mu/(\mu - 1)$. The symmetries in the general case can also be realized as rotational symmetries of the sphere similarly to the two cases considered above. By a suitable projective transformation the four points, A, B, C, D can be transformed into the vertices of a tetrahedron which has opposite edges equal in pairs (a disphenoid). Such a tetrahedron has three orthogonal dyad axes of symmetry. These axes are the joins of the midpoints of opposite edges. If the cross-ratio is real, the tetrahedron is flattened into a rectangle but the three symmetry axes remain.

To see that such a transformation exists consider the three pairs (E,F), (G,H), (K,L) of united points for the three involutions which send (A,B,C,D) into (B,A,D,C), (C,D,A,B) and (D,C,B,A), respectively. Now the involution which sends (A,B,C,D) into (B,A,D,C) transforms the other involutions into themselves. It therefore sends G into H and K into L. Hence, (E,F) is harmonic with respect to (G,H) and also with respect to (K,L). Similarly (G,H) is harmonic with respect to (K,L). Now, E, G, F, H can be transformed (as above) into four points equally spaced, in that order, around the equator. K and L will then be the north and south poles, so that the six points form the

vertices of a regular octahedron. The three involutions are then represented as rotations through π about the three axes EF, GH, KL. The point A is rotated into B, C, D by means of these involutions giving the symmetrical tetrahedron described above.

This symmetrical representation of the points A, B, C, D is of interest because it is related to Petrov's canonical representation of the Riemann tensor with $R_{\mu\nu} = 0$ (*17*, *20*). The rest frame in which the gravitational principal null directions appear to have this symmetrical form determines the canonical time axis, the three canonical space axes arising from the three axes of symmetry. These four axes are orthogonal to each other and are called the Riemann principal directions. They are uniquely defined provided that A, B, C, D are all distinct. If A, B, C, D coincide in pairs they can still be considered to exist but they are not uniquely defined.

The rotational symmetries of the tetrahedron A B C D in the general case give rise to the corresponding symmetries for $R_{\mu\nu\rho\sigma}$, since being dyad axes the only other possibility would be $R_{\mu\nu\rho\sigma} \rightarrow -R_{\mu\nu\rho\sigma}$ (a duality rotation of π).[9] Such an alternative is easily ruled out as impossible. It follows that, for the canonical choice of axes,

$$R_{ijkl} = 0 \quad \text{whenever} \quad i = k \quad \text{and} \quad j \neq l$$

as is required in Petrov's canonical form. Conversely, the above condition is sufficient for the Riemann principal directions to be the axes.

The usual definitions of the Riemann principal directions is in terms of the intersections of certain planes which are determined by the "eigenbivectors" of $R_{\mu\nu\rho\sigma}$, i.e., from the nonzero (complex) skew tensors $x^{\mu\nu}$ which satisfy a relation

$$R^{\mu\nu}{}_{\rho\sigma}x^{\rho\sigma} = \alpha x^{\mu\nu}. \tag{4.7}$$

Writing this in a spinor form with

$$x^{AC'BD'} = \tfrac{1}{2}\{\eta^{AB}\epsilon^{C'D'} + \epsilon^{AB}\bar{\zeta}^{C'D'}\}$$

(see 1.3) η^{AB} and ζ^{AB} being symmetric, (4.7) becomes

$$\psi^{AB}{}_{EF}\eta^{EF}\epsilon^{C'D'} + \epsilon^{AB}\bar{\psi}^{C'D'}{}_{E'F'}\bar{\zeta}^{E'F'} = \alpha\{\eta^{AB}\epsilon^{C'D'} + \epsilon^{AB}\bar{\zeta}^{C'D'}\}$$

(since $\phi_{ABC'D'} = 0$, $\lambda = 0$) so that

$$\psi^{AB}{}_{EF}\eta^{EF} = \alpha\eta^{AB}, \quad \psi^{AB}{}_{EF}\zeta^{EF} = \bar{\alpha}\zeta^{AB}.$$

One or the other of η^{AB}, ζ^{AB} may be zero. The eigenbivectors of $R^{\mu\nu}{}_{\rho\sigma}$ are thus

[9] In the special cases where the set of points A, B, C, D has an additional rotational symmetry, this does not always lead to a corresponding symmetry of $R_{\mu\nu\rho\sigma}$, although it does for the case when A, B, C, D coincide in pairs. In particular, in the equianharmonic case, the triad axes of symmetry give rise to duality rotations through angles $2\pi/3$, $4\pi/3$.

expressible in terms of "eigenspinors" of $\psi^{AB}{}_{CD}$, the eigenvalues of $R^{\mu\nu}{}_{\rho\sigma}$ being being those of $\psi^{AB}{}_{CD}$ and their complex conjugates. Witten (*6*) also considers these eigenspinors.

Now if the eigenvalues of $\psi^{AB}{}_{CD}$ are α_1, α_2, α_3 (the space of symmetric ξ^{AB} being three-dimensional) we have

$$\alpha_1 + \alpha_2 + \alpha_3 = \psi^{AB}{}_{AB} = 0,$$
$$\alpha_1^2 + \alpha_2^2 + \alpha_3^2 = \psi^{AB}{}_{CD}\psi^{CD}{}_{AB} = I,$$
$$\alpha_1^3 + \alpha_2^3 + \alpha_3^3 = \psi^{AB}{}_{CD}\psi^{CD}{}_{EF}\psi^{EF}{}_{AB} = J.$$

With the expressions for I and J given in (4.6), it is easily verified that these relations are satisfied by

$$\alpha_1 = \kappa(2\mu - 1), \quad \alpha_2 = \kappa(2 - \mu), \quad \alpha_3 = \kappa(-1 - \mu). \tag{4.8}$$

The six eigenvalues of $R^{\mu\nu}{}_{\rho\sigma}$ are therefore these three numbers and their complex conjugates. It will be seen that the vanishing of just one of the eigenvalues (4.8) is the condition for the principal null directions to form a harmonic set. If two of them vanish they must all vanish and $I = J = 0$. This is the condition for at least three of the principal null directions to coincide (since they form both a harmonic and an equianharmonic set). If two of the eigenvalues (4.8) coincide, this is the condition $\mu = 0$, 1, or ∞ for a pair of principal null directions to coincide. This is the case $I^3 = 6J^2$ (*18*, p. 198).

The three eigenspinors η^{AB}, ζ^{AB}, θ^{AB} of $\psi^{AB}{}_{CD}$ will next be considered. They are symmetric and therefore each is expressible as a symmetrized product of a pair of one-index spinors (see 4.1). Each of η^{AB}, ζ^{AB}, θ^{AB} corresponds to a pair of points on the projective line considered earlier, so in the general case we have six points on this line determined by A, B, C, D. These can only be E, F, G, H, K, L since a general quartic form has only one sextic covariant (*18*, pp. 92, 94). This sextic covariant is

$$\psi_{PQRA}\psi^{PQ}{}_{BC}\psi^{R}{}_{DEF}\xi^{A}\xi^{B}\xi^{C}\xi^{D}\xi^{E}\xi^{F},$$

whence

$$\psi_{PQR(A}\psi^{PQ}{}_{BC}\psi^{R}{}_{DEF)} = \eta_{(AB}\zeta_{CD}\theta_{EF)}$$

choosing the scale factor suitably. The vanishing of this expression is the condition for A, B, C, D to coincide in pairs, since E, F, G, H, K, L are not then defined uniquely. It does not vanish if just two of A, B, C, D coincide, or if they coincide three and one.

The planes determined by the eigenbivectors of $R^{\mu\nu}{}_{\rho\sigma}$ are those determined by η^{AB}, ζ^{AB}, θ^{AB}. They are therefore the three planes of the pairs of null directions

corresponding to EF, GH, KL and the three orthogonal complements of these planes. Their intersections give the Riemann principal directions defined here, as is required. This is easily seen from the symmetrical representation of A, B, C, D given above.

These considerations have so far been essentially only concerned with Petrov's tensors $R_{\mu\nu\rho\sigma}$ of Type I. This is the case when the eigenbivectors of $R^{\mu\nu}{}_{\rho\sigma}$ span the six-dimensional space of bivectors. In special cases these eigenbivectors span only a four-dimensional space (Type II) and in very special cases, a two-dimensional space (Type III). In spinor terms, this means that Type I occurs when the eigenspinors of $\psi^{AB}{}_{CD}$ span a three-dimensional space, Type II when they span a two-dimensional space and Type III when they span only a one-dimensional space. Thus, Type II can only occur when at least two of the eigenvalues (4.8) are equal and Type III when they are all equal (and therefore all zero). We have seen that equality of eigenvalues implies coincidences among A, B, C, D so the cases where such coincidences occur must now be considered.

There are six different cases to be distinguished including the general case [1111] where the null directions are all distinct. There is the case [211] where exactly two of them coincide, [22] where they coincide in pairs, [31] where they coincide three and one, and [4] where all four directions are the same. Finally, there is the case [—] when $\psi_{ABCD} = 0$ and the null directions are undefined. This gives us a natural classification of Riemann tensors in empty space into six types (see also Géhéniau (*21*) for a closely related procedure[10]). In each case, the eigenspinors can be obtained by observing what happens to E, F, G, H, K, L when A, B, C, D are specialized. However, this must be done with care so that possible limiting positions of E, F, G, H, K, L are not omitted. Figure 1 shows how the different special cases arise from one another. The vertical specializations can be carried out keeping the positions of E, F, G, H, K, L fixed, but in the diagonal specializations, further pairs of them are forced to coincide. (For example, in the case [1111] → [211] if B → X and A → X, we have (G, H → (X, X), (K, L) → (X, X) and (E, F) → (X, Y) where Y is the harmonic conjugate of X with respect to the limiting positions of C and D.) The Petrov type for each case may be obtained in this way and the results are shown in Fig. 1. Each column corresponds to a particular type. Thus, [1111], [22], and [—] are Type I, [211] and [4] are Type II, while [31] is Type III. The different rows can be distinguished by the invariants I and J (or by the eigenvalues). Hence the invariants and Petrov type together serve to characterize ψ_{ABCD}.

It is of interest to see how this classification is in accord with that given by the classical canonical form of $\psi^{AB}{}_{CD}$ considered as a (3×3) matrix. These corresponding canonical forms are given in Fig. 2.

The various algebraic conditions for each case (or one of its specializations)

[10] *Note added in proof:* See also, more explicitly, Debever (*15*, *16*).

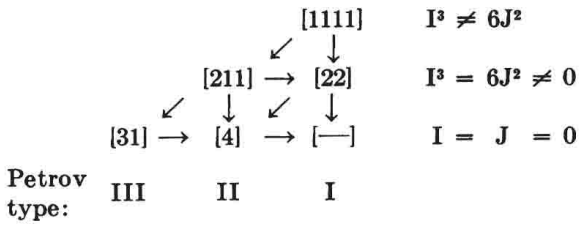


FIG. 1. Classification scheme for ψ_{ABCD} in terms of coincidences between principal null directions.

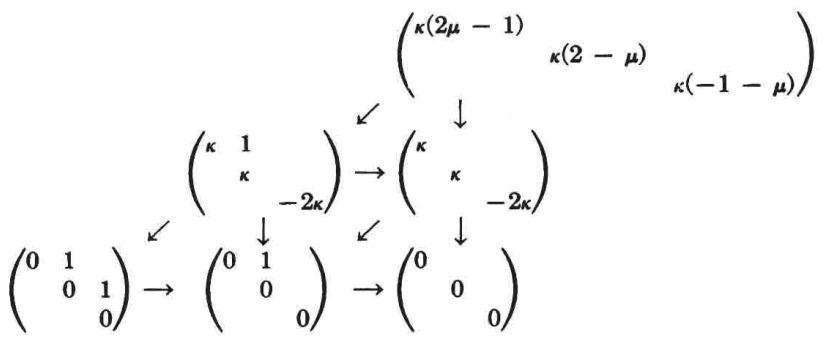

$\begin{pmatrix} \kappa(2\mu-1) & & \\ & \kappa(2-\mu) & \\ & & \kappa(-1-\mu) \end{pmatrix}$
$\begin{pmatrix} \kappa & 1 & \\ & \kappa & \\ & & -2\kappa \end{pmatrix} \rightarrow \begin{pmatrix} \kappa & & \\ & \kappa & \\ & & -2\kappa \end{pmatrix}$
$\begin{pmatrix} 0 & 1 & \\ & 0 & 1 \\ & & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 1 & \\ & 0 & \\ & & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & & \\ & 0 & \\ & & 0 \end{pmatrix}$

FIG. 2. Classification in terms of matrix canonical form of $\psi^{AB}{}_{CD}$.

to occur may be collected together as follows:

$$[211]: I^3 = 6J^2, \quad [22]: \psi_{PQR(A}\psi^{PQ}{}_{BC}\psi^{R}{}_{DEF)} = 0, \quad [31]: I = J = 0,$$

$$[4]: \psi_{(AB}{}^{EF}\psi_{CD)EF} = 0, \quad [—]: \psi_{ABCD} = 0.$$

The only case that has not already been dealt with is the condition for [4] to occur. The quartic form $\psi_{AB}{}^{EF}\psi_{CDEF}\xi^A\xi^B\xi^C\xi^D$ is the Hessian of the form

$$\psi_{ABCD}\xi^A\xi^B\xi^C\xi^D$$

and its vanishing is known to be the condition for the latter form to be a perfect fourth power (*18*, p. 235). The interest of this condition lies in the fact that $\psi_{(AB}{}^{EF}\psi_{CD)EF}$ is precisely the term (in the case $\lambda = 0$) which prevents Eq. (3.8) from being a covariant wave equation[11] for ψ_{ABCD}. Thus, plane wave solutions can only reasonably be expected in case [4].[12] This is Petrov's Type II with vanishing invariants and is apparently the case characteristic of a "*pure*" gravitational radiation field (*8*, *22*, and *23*). The other cases which might conceivably also be considered as "pure gravitational radiation" are [211] and [31] (see *17*).

[11] Case [4] therefore appears to be the only case (apart from [—]) in which the gravitational field has no "gravitational mass". See also Bondi *et al.* (*23*), p. 532.

[12] However, a point perhaps worth mentioning is that in case [22], ψ_{ABCD} and $\psi_{(AB}{}^{EF}\psi_{CD)EF}$ are proportional.

Case [211] would seem to be wrong since [22], which is a special case of it, would also have to be considered as pure gravitational radiation. But we have seen that the Schwarzschild solution is [22].

Case [31] is, however, worthy of consideration in this respect since it shares with Case [4] the property that the gravitational density (4.5) can be made as small as we please by a suitable choice of time axis ("following the wave"). If

$$\psi_{ABCD} = \alpha_{(A}\alpha_B\alpha_C\beta_{D)}$$

and

$$l_\mu = a_\mu + \epsilon x_\mu , \quad a_{AB'} = \alpha_A\bar{\alpha}_{B'} , \quad b_{AB'} = \beta_A\bar{\beta}_{B'} ,$$

where $\epsilon > 0$ is small and x_μ is time-like pointing to the future, we have

$$\frac{T_{\mu\nu\rho\sigma}l^\mu l^\nu l^\rho l^\sigma}{(l_\tau l^\tau)^2} \doteq \frac{\frac{1}{4}\epsilon^3 \mid \beta_A\alpha^A \mid^2 (\alpha_B\bar{\alpha}_{C'}x^{BC'})^3}{4\epsilon^2(a_\tau x^\tau)^2} ,$$

$$= \epsilon\frac{(b_\mu a^\mu)(a_\nu x^\nu)^3}{16(a_\tau x^\tau)^2} .$$

If $\beta_A = \alpha_A$, the right-hand side would be of order ϵ^2 instead of ϵ. Thus, the gravitational density tends to zero for observers, whose velocity approaches the multiple principal null direction, both in Case [31] and in Case [4], but it tends to zero more rapidly in Case [4]. It would appear to be correct to call Case [4] "pure" radiation field[13] but not Case [31]. Case [4] is like a null electromagnetic field ("pure" electromagnetic radiation field) in that it determines only one null direction, and in that it is the general limiting case obtained as a result of a high-velocity Lorentz transformation (see also *24*). (However, it is worth remarking that for a null electromagnetic field, the energy $T_{\mu\nu}l^\mu l^\nu/(l_\tau l^\tau)$ can only be made to tend to zero to order ϵ by "following the wave", like Case [31] above.)

The invariants of ψ_{ABCD} have been treated in considerable detail above. It now remains to give a brief discussion of the combined system ψ_{ABCD} , ϕ_{AB} for the case when electromagnetic field is present. We expect to find just three more complex invariants, since ϕ_{AB} is determined by its phase and magnitude, and by the positions relative to A, B, C, D of the two complex points Y and Z on the argand sphere, corresponding to the electromagnetic principal null directions. There is the obvious invariant

$$K = \phi_{AB}\phi^{AB}$$

of ϕ_{AB} alone. This is the discriminant of the binary form $\phi_{AB}\xi^A\xi^B$, the condition

[13] It is probably preferable, however, to call case [4] simply a *null* gravitational field (as suggested by Robinson) analogously to the electromagnetic case.

K = 0 being necessary and sufficient for the points X and Y to coincide, that is, for the field to be null.[13a] The list is completed by the two independent invariants

$$L = \phi_{AB}\psi^{AB}{}_{CD}\phi^{CD}, \quad M = \phi_{AB}\psi^{AB}{}_{CD}\psi^{CD}{}_{EF}\phi^{EF}.$$

The fact that I, J, K, L, M are in general independent is most easily seen if $\psi^{AB}{}_{CD}$ is thought of as a matrix and ϕ^{AB} as a "vector" which may then be expanded in terms of the eigenspinors of $\psi^{AB}{}_{CD}$ with arbitrary coefficients. K, L, and M then become independent linear functions of the squares of these coefficients.

However, I, J, K, L, and M do not form a complete system of invariants in the sense of invariant theory (*18*). That is, not every algebraic invariant of ψ_{ABCD} and ϕ_{AB} can be expressed as a polynomial in I, ... , M. The invariant

$$N = \phi_{AB}\psi^{AB}{}_{CD}\psi^{CD}{}_{EF}\phi^{E}{}_{G}\psi^{FG}{}_{PQ}\phi^{PQ}$$

clearly is not even a rational function of I, ... , M since every such functions is of even order in ϕ_{AB} . Also N does not vanish identically. On the other hand, N is *algebraically* dependent on I, ... , M, there being the *syzygy*

$$N^2 = \tfrac{1}{2}JKLM - \tfrac{1}{6}JL^3 - \tfrac{1}{2}M^3 - \tfrac{1}{8}I^2KL^2 - \tfrac{1}{6}IJK^2L - \tfrac{1}{18}J^2K^3 + \tfrac{1}{4}IKM^2 + \tfrac{1}{4}IL^2M .$$

The system I, J, K, L. M, N does, in fact, form a complete system of invariants for ψ_{ABCD} and ϕ_{AB} .

The condition for an electromagnetic principal null direction to coincide with a gravitational principal null direction is that the resultant of the quartic and quadratic forms should vanish. Expressed in terms of invariants this condition turns out to be

$$2K^2I - 4KM + L^2 = 0.$$

The condition for both electromagnetic null directions to lie along a gravitational null direction is therefore

$$K = 0, \quad L = 0.$$

The electromagnetic and gravitational fields together have ten independent real invariants, namely the real and imaginary parts of I, J, K, L, M. However, only nine of these are determined by the curvature $R_{\mu\nu\rho\sigma}$ since it is unaffected by duality rotations of the electromagnetic field. These are the nine independent

[13a] The real and imaginary parts of K are the usual invariants $F_{\mu\nu}F^{\mu\nu}$ and $1/2\sqrt{-g}\,F^{\mu\nu}F^{\rho\sigma}\epsilon_{\mu\nu\rho\sigma}$ respectively, of $F_{\mu\nu}$.

real invariants of ψ_{ABCD} and $\phi_{ABC'D'} = \phi_{AB}\bar{\phi}_{C'D'}$. The phase of ϕ_{AB} is undetermined by $\phi_{ABC'D'}$, so we can take for these invariants[14]

$$\mathrm{I},\ \mathrm{J},\ |\ \mathrm{K}\ |,\ |\ \mathrm{L}\ |,\ |\ \mathrm{M}\ |$$

and the arguments of the two ratios

$$\mathrm{K{:}L{:}M}.$$

(The invariants $|\ \mathrm{K}\ |^2$, $|\ \mathrm{L}\ |^2$, $|\ \mathrm{M}\ |^2$, $\mathrm{K}\bar{\mathrm{L}}$, $\mathrm{L}\bar{\mathrm{M}}$, $\mathrm{M}\bar{\mathrm{K}}$ are easily expressible in terms of ψ_{ABCD} and $\phi_{ABC'D'}$.)

5. ANALYTIC SOLUTIONS OF EINSTEIN'S EQUATIONS

Let $\mathfrak{M}$ be an analytic (connected) Riemannian manifold. Then starting from any point O on $\mathfrak{M}$ at which the curvature tensor $\mathrm{R}_{\mu\nu\rho\sigma}$ and all its covariant derivatives are known, it is possible to calculate the curvature tensor (and its derivatives) at any other point by means of a power series:

$$(\mathrm{R}_{\mu\nu\rho\sigma})_x = (\mathrm{R}_{\mu\nu\rho\sigma})_0 + x^\alpha(\partial_\alpha \mathrm{R}_{\mu\nu\rho\sigma})_0 + \frac{1}{2!}x^\alpha x^\beta(\partial_\alpha\partial_\beta \mathrm{R}_{\mu\nu\rho\sigma})_0 + \cdots. \tag{5.1}$$

The point x is that point on $\mathfrak{M}$ whose geodesic distance from O is $\sqrt{(x_\alpha x^\alpha)}$ and which lies on the geodesic through O which starts off in the direction of x^α(Riemannian coordinates). (If x^α is null this has to be interpreted suitably.) The $\mathrm{R}_{\mu\nu\rho\sigma}$ at the point x is referred to axes which are those at O transferred in parallel along this geodesic. If the power series does not converge, the point x may be reached in several steps, using intermediate points, in the manner of analytic continuation. This power series expression and its convergence is considered by Thomas (*25*, p. 234).

Equation (5.1) is a special case of the more general situation, whereby any analytic tensor field may be calculated from a knowledge of the tensor and all its covariant derivatives at the point 0 alone:

$$\begin{aligned}(f\ldots)_x &= (f\ldots)_0 + x^\alpha(\partial_\alpha f\ldots)_0 + \frac{1}{2!}x^\alpha x^\beta(\partial_\alpha\partial_\beta f\ldots)_0 + \cdots \\ &= [\exp(x^\alpha\partial_\alpha)f\ldots]_0 = \left[\lim_{n\to\infty}\left(1+\frac{1}{n}x^\alpha\partial_\alpha\right)^n f\ldots\right]_0.\end{aligned} \tag{5.2}$$

[14] When the electromagnetic field is null there still remain the seven real invariants given by I, J, | L |, | M | and the argument of L/M. Thus Witten (*6*) is mistaken when he claims that there remain only the four real invariants of ψ_{ABCD} in this case. For example, the invariant $\phi_{ABE'F'}\psi^{AB}{}_{CD}\bar{\psi}^{E'F'}{}_{G'H'}\phi^{CDG'H'} = |\ \mathrm{L}\ |^2$ need not vanish when K = 0. Such an invariant could appear as a quotient of invariants built up from Witten's list.

The ∂_α's are to be taken as acting only on $f\cdots$ and not on x^α. (This last expression can be used to obtain the power series expression since

$$(f\ldots)_{\epsilon x} = [(1 + \epsilon x^\alpha \partial_\alpha) f\ldots]_0 + O(\epsilon^2),$$

which may be applied n times with $\epsilon = 1/n$, giving $(f\ldots)_x$ correct to order $1/n$).

These power series can be used as the basis for a coordinate-free approach to Riemannian geometry. Instead of specifying a space by giving the metric tensor $g_{\mu\nu}$ as a function of some coordinates, the space may be determined (except possibly for some of its topological properties in the large) by specifying $R_{\mu\nu\rho\sigma}$, $\partial_\alpha R_{\mu\nu\rho\sigma}$, $\partial_\alpha\partial_\beta R_{\mu\nu\rho\sigma}$, $\cdots$ at a point O. To specify a set of tensors at a point does not require coordinates since their algebraic tensorial properties need only be given. The metric tensors $g_{\mu\nu}$, $g^{\mu\nu}$ and the alternating tensor $\sqrt{\pm g}\epsilon_{\mu\ldots\sigma}$ are also supposed to be specified at the point O. They are an essential part of the tensor algebra at O.

A difficulty about specifying a space in this way is that $R_{\mu\nu\rho\sigma}$, $\partial_\alpha R_{\mu\nu\rho\sigma}$, $\partial_\alpha\partial_\beta R_{\mu\nu\rho\sigma}$, $\cdots$ are not algebraically (tensorially) independent of one another. Relation (2.1) implies identities (Ricci) connecting second derivatives with the curvature tensor, and also there is the Bianchi identity which is the consistency condition for (2.1). The Bianchi identity is in fact the only consistency condition required (*25*, pp. 131, 132). Applying these two types of identity to the higher derivatives of $R_{\mu\nu\rho\sigma}$ a host of relations is obtained. It is therefore of importance to be able to single out a set of tensors which are algebraically independent (in the general case) and from which $R_{\mu\nu\rho\sigma}$ and all its derivatives are obtainable by algebraic operations. It is possible to show that the following set of tensors, in fact, has all these properties:

$$Q^{\mu\nu}{}_{\rho\sigma} = R^{\mu}{}_{(\rho}{}^{\nu}{}_{\sigma)}, \quad Q^{\mu\nu}{}_{\rho\sigma\alpha} = \partial_{(\alpha} R^{\mu}{}_{\rho}{}^{\nu}{}_{\sigma)}, \quad Q^{\mu\nu}{}_{\rho\sigma\alpha\beta} = \partial_{(\alpha}\partial_\beta R^{\mu}{}_{\rho}{}^{\nu}{}_{\sigma)}, \quad \text{etc.}$$

Each $Q\ldots$ has the symmetry given by a Young tableau operator corresponding to a partition $(r - 2, 2)$. That is to say, we have

$$Q_{\mu\nu\rho\sigma\cdots\beta} = Q_{(\mu\nu)(\rho\sigma\cdots\beta)} \quad \text{and} \quad Q_{\mu(\nu\rho\sigma\cdots\beta)} = 0.$$

Apart from these symmetries and from certain considerations of convergence, the Q's may be chosen arbitrarily.[15] Unfortunately, however, if it is required to impose a condition such as Einstein's $R^{\mu}{}_{\nu\mu\sigma} = 0$ (or $= \lambda g_{\nu\sigma}$) on the space, this implies a condition not only on $Q^{\mu\nu}{}_{\rho\sigma}$, but also on $Q^{\mu\nu}{}_{\rho\sigma\alpha}$, $Q^{\mu\nu}{}_{\rho\sigma\alpha\beta}$, etc. These conditions are all linear, but they appear to be somewhat complicated. It seems for this reason that an approach based explicitly on these Q's would not be usually very convenient for general relativity. (However, in a later paper it is proposed

[15] These Q's are somewhat analogous to (but different from) the "normal tensors" (see Thomas *25*, p. 102).

to give a class of special solutions using this method.) On the other hand, if a spinor approach is used, these linear conditions take on a particularly simple form. This approach will now be described in more detail.

Suppose that $\mathfrak{M}$ has four dimensions and signature $(+---)$, and that $R^{\mu}{}_{\nu\mu\sigma} = \lambda g_{\nu\sigma}$. Then we have seen that $R_{\mu\nu\rho\sigma}$ can be represented uniquely by a totally symmetric spinor ψ_{ABCD} (λ being known). We wish to find a set of algebraically independent spinors from which

$$\psi_{ABCD}\,, \quad \partial_E{}^{P'}\psi_{ABCD}\,, \quad \partial_E{}^{P'}\partial_F{}^{Q'}\psi_{ABCD}\,, \cdots \tag{5.3}$$

(at the point O) can be constructed by means of algebraic spinor operations. The identities relating the spinors (5.3) arise from the equivalent of the Bianchi identity, namely (3.5):

$$\partial^{AP'}\psi_{ABCD} = 0 \quad \text{or} \quad \epsilon^{EA}\partial_E{}^{P'}\psi_{ABCD} = 0 \tag{5.4}$$

and the equivalent of (2.1), namely, (2.14), (2.15), (2.16), and (2.17):

$$\epsilon_{R'S'}\{\partial_G{}^{R'}\partial_H{}^{S'} + \partial_H{}^{R'}\partial_G{}^{S'}\}\xi_A = \psi_{GHAB}\xi^B - \frac{\lambda}{3}\{\xi_G\epsilon_{HA} + \xi_H\epsilon_{GA}\}, \tag{5.5}$$

$$\epsilon_{R'S'}\{\partial_G{}^{R'}\partial_H{}^{S'} + \partial_H{}^{R'}\partial_G{}^{S'}\}\eta^{P'} = 0$$

$$\epsilon^{GH}\{\partial_G{}^{R'}\partial_H{}^{S'} + \partial_G{}^{S'}\partial_H{}^{R'}\}\xi_A = 0,$$

$$\epsilon^{GH}\{\partial_G{}^{R'}\partial_H{}^{S'} + \partial_G{}^{S'}\partial_H{}^{R'}\}\eta^{P'} = \bar{\psi}^{R'S'P'}{}_{Q'}\eta^{Q'} - \frac{\lambda}{3}\{\eta^{R'}\epsilon^{S'P'} + \eta^{S'}\epsilon^{R'P'}\} \tag{5.6}$$

(see 3.7 and 3.3) applied to ψ_{ABCD} and its derivatives.

The various derivatives of (5.4) must all hold identically also. Hence the algebraic relations on the spinors (5.3) arising from (5.4) are

$$\epsilon^{HA}(\partial_E{}^{P'} \cdots \partial_G{}^{R'}\partial_H{}^{S'}\psi_{ABCD}) = 0. \tag{5.7}$$

This expresses a condition on (namely, the vanishing of) the part of $\partial_E{}^{P'} \cdots \partial_H{}^{S'}\psi_{ABCD}$ which is skew in H, A and says nothing about the part symmetric in H, A. Moreover the relations (5.5) connect

$$\epsilon_{R'S'}(\partial_E{}^{P'} \cdots \partial_G{}^{R'}\partial_H{}^{S'} \cdots \partial_K{}^{V'}\psi_{ABCD}) + \epsilon_{R'S'}(\partial_E{}^{P'} \cdots \partial_H{}^{R'}\partial_G{}^{S'} \cdots \partial_K{}^{V'}\psi_{ABCD})$$

with lower derivatives of ψ_{ABCD}, while (5.6) connect

$$\epsilon^{GH}(\partial_E{}^{P'} \cdots \partial_G{}^{R'}\partial_H{}^{S'} \cdots \partial_K{}^{V'}\psi_{ABCD}) + \epsilon^{GH}(\partial_E{}^{P'} \cdots \partial_G{}^{S'}\partial_H{}^{R'} \cdots \partial_K{}^{V'}\psi_{ABCD'})$$

with lower derivatives of ψ_{ABCD}. These express conditions only on parts of $\partial_E{}^{P'} \cdots \partial_K{}^{V'}\psi_{ABCD}$ which are skew in a pair of primed indices or in a pair of unprimed indices. Thus the algebraic relations arising from (5.4), (5.5), and (5.6) connecting the spinors (5.3) are all concerned with parts of $\partial_E{}^{P'} \cdots \partial_K{}^{V'}\psi_{ABCD}$

which are skew in at least one pair of indices. They imply no conditions on the parts totally symmetric in all primed indices and in all unprimed indices. (It might, perhaps, be thought that other relations could be obtained by expanding skew parts of $\partial_E{}^{P'} \cdots \partial_K{}^{V'} \psi_{ABCD}$ in two different ways. However, these all lead back to (5.7) which is the only consistency condition implied.) Hence the spinors

$$\psi_{ABCD}\,,\ \psi_{ABCDE}{}^{P'} = \partial_{(E}{}^{P'} \psi_{ABCD)}\,,\ \psi_{ABCDEF}{}^{P'Q'} = \partial_{(E}{}^{(P'} \partial_F{}^{Q')} \psi_{ABCD)}\,,\ \cdots \tag{5.8}$$

are all algebraically independent and can therefore be specified arbitrarily (apart from convergence considerations) at the point 0.

The problem is now to show, conversely, that all the spinors (5.3) can be obtained algebraically from the spinors (5.8). For then ψ_{ABCD}, ψ_{ABCDEP}, $\psi_{ABCDEFP'Q'}$, $\cdots$ will be a complete set of algebraically independent spinors at O, which can be used to generate the space $\mathfrak{M}$. In order to show that they form such a complete set, an argument by induction will be used. We wish to express $\partial_E{}^{P'} \cdots \partial_K{}^{V'} \psi_{ABCD}$ in terms of $\psi_{ABCDE\cdots K}{}^{P'\cdots V'}$ and lower order derivatives of ψ_{ABCD} since it may be supposed as the inductive hypothesis that all these lower derivatives have already been expressed algebraically in terms of symmetrized derivatives $\psi_{AB\cdots G}{}^{P'\cdots R'}$. Now, if we add together all the spinors obtained from $\partial_E{}^{P'} \cdots \partial_K{}^{V'} \psi_{ABCD}$ by permuting P′, $\cdots$, V′ in all possible ways and A, B, C, D, E, $\cdots$, K in all possible ways, we get a multiple of $\psi_{AB\cdots K}{}^{P'\cdots V'}$. Thus, if it can be shown that each of the spinors obtained by such permutations differs from $\partial_E{}^{P'} \cdots \partial_K{}^{V'} \psi_{ABCD}$ by expressions involving only lower derivatives of ψ_{ABCD} the result will be proved. The spinor $\partial_E{}^{P'} \cdots \partial_K{}^{V'} \psi_{ABCD}$ will then be seen to differ from $\psi_{AB\cdots K}{}^{P'\cdots V'}$ by a spinor built up from lower derivatives of ψ_{ABCD}.

Any two spinors obtained by such a permutation of indices from

$$\partial_E{}^{P'} \cdots \partial_K{}^{V'} \psi_{ABCD}$$

will be called *equivalent* (denoted by $\sim$) if they differ from each other by expressions built up from lower order derivatives of ψ_{ABCD}. This is clearly an equivalence relation. It is required to show that all such spinors are, in fact, equivalent to one another. Now since

$$\begin{aligned}\partial_W{}^{X'} \partial_Y{}^{Z'} - \partial_Y{}^{Z'} \partial_W{}^{X'} \equiv {}& \tfrac{1}{2} \epsilon^{X'Z'} \epsilon_{M'N'} \{\partial_W{}^{M'} \partial_Y{}^{N'} + \partial_Y{}^{M'} \partial_W{}^{N'}\} \\ & + \tfrac{1}{2} \epsilon_{WY} \epsilon^{ST} \{\partial_S{}^{X'} \partial_T{}^{Z'} + \partial_S{}^{Z'} \partial_T{}^{X'}\}\end{aligned}$$

(see 2.12), we have, applying (5.5) and (5.6)

$$\cdots \partial_W{}^{X'} \partial_Y{}^{Z'} \cdots \psi_{ABCD} \sim \cdots \partial_Y{}^{Z'} \partial_W{}^{X'} \cdots \psi_{ABCD}\,.$$

Hence any permutation of the $\partial_M{}^{N'}$ symbols gives rise to an equivalent spinor. (Any permutation can be expressed as a product of transpositions of adjacent elements.) That is, any permutation of P′, $\cdots$, V′ can be applied to

$\partial_E{}^{P'} \cdots \partial_K{}^{V'} \psi_{ABCD}$ provided that the same permutation is applied to E, $\cdots$, K and an equivalent spinor is obtained. It remains to show that E, $\cdots$, K, A, B, C, D can be permuted independently and an equivalent spinor is still obtained. The symmetry of ψ_{ABCD} implies that A, B, C, D can be permuted without change. Furthermore, from 5.7, K and A can be interchanged in $\partial_E{}^{P'} \cdots \partial_K{}^{V'} \psi_{ABCD}$. Also,

$$\cdots \partial_Y{}^{Z'} \cdots \partial_K{}^{V'} \psi_{ABCD} \sim \cdots \partial_K{}^{V'} \cdots \partial_Y{}^{Z'} \psi_{ABCD},$$
$$\sim \cdots \partial_K{}^{V'} \cdots \partial_A{}^{Z'} \psi_{YBCD} \sim \cdots \partial_A{}^{Z'} \cdots \partial_K{}^{V'} \psi_{YBCD}$$

so that A can be interchanged with any other unprimed index and an equivalent spinor is obtained. It follows that any pair of unprimed indices can be interchanged since

$$\cdots \partial_W{}^{X'} \cdots \partial_Y{}^{Z'} \cdots \psi_{ABCD} \sim \cdots \partial_W{}^{X'} \cdots \partial_A{}^{Z'} \cdots \psi_{YBCD},$$
$$\sim \cdots \partial_Y{}^{X'} \cdots \partial_A{}^{Z'} \cdots \psi_{WBCD} \sim \cdots \partial_Y{}^{X'} \cdots \partial_W{}^{Z'} \cdots \psi_{ABCD}.$$

Hence all the spinors are equivalent and the result is proved.

As examples of the above, we have

$$\partial_E{}^{P'} \psi_{ABCD} = \psi_{ABCDE}{}^{P'},$$
$$\partial_E{}^{P'} \partial_F{}^{Q'} \psi_{ABCD} = \psi_{ABCDEF}{}^{P'Q'} + \epsilon_{EF} \epsilon^{P'Q'} \{\tfrac{3}{4} \psi_{(AB}{}^{GH} \psi_{CD)GH} - \tfrac{1}{2} \lambda \psi_{ABCD}\}$$
$$+ \epsilon^{P'Q'} \{\psi_{(ABC}{}^{G} \psi_{D)EFG} + \tfrac{1}{3} \lambda \psi_{E(ABC} \epsilon_{D)F} + \tfrac{1}{3} \lambda \psi_{F(ABC} \epsilon_{D)E}\}.$$

Higher derivatives involve $\bar{\psi}_{A'B'C'D'}$, $\bar{\psi}_{A'B'C'D'E'P}$, $\cdots$ also. We have from (5.1), with $\psi_{ABCD} = (\psi_{ABCD})_0$, etc.,

$$(\psi_{ABCD})_x = \psi_{ABCD} + x^{EP'} \partial_{EP'} \psi_{ABCD} + \frac{1}{2!} x^{EP'} x^{FQ'} \partial_{EP'} \partial_{FQ'} \psi_{ABCD} + \cdots.$$

Hence

$$(\psi_{ABCD})_x = \psi_{ABCD} + x^{EP'} \psi_{ABCDEP'} + \tfrac{1}{2} x^{EP'} x^{FQ'} \psi_{ABCDEFP'Q'}$$
$$+ \tfrac{1}{2} (x_{EP'} x^{EP'}) \{\tfrac{3}{4} \psi_{(AB}{}^{GH} \psi_{CD)GH} - \tfrac{1}{2} \lambda \psi_{ABCD}\} + O(x^3).$$

It is possible to obtain a class of exact solutions for gravitational plane waves using this method. Such solutions, obtained using more conventional methods, have been known for some time (for references, see Bondi *et al.*, *23*). Let

$$\psi_{ABCD} = \alpha_0 \pi_A \pi_B \pi_C \pi_D, \quad \psi_{ABCDEP'} = \alpha_1 \pi_A \pi_B \pi_C \pi_D \pi_E \bar{\pi}_{P'},$$
$$\psi_{ABCDEFP'Q'} = \alpha_2 \pi_A \cdots \pi_F \bar{\pi}_{P'} \bar{\pi}_{Q'}, \cdots \tag{5.9}$$

at the point O, where π_A is a spinor corresponding to the null direction giving the direction of motion of the wave and α_0, α_1, $\cdots$ are complex numbers. Suppose

$\lambda = 0$. It will now be shown that the unsymmetrized derivatives of ψ_{ABCD} are all equal to the symmetrized derivatives, so the situation is much simplified in this case. As an inductive hypothesis we assume that all the derivatives of ψ_{ABCD} of lower order than $\partial_E{}^{P'} \cdots \partial_K{}^{V'}\psi_{ABCD}$ are already symmetric and therefore equal to the corresponding expressions 5.9. The argument given above shows that $\partial_E{}^{P'} \cdots \partial_K{}^{V'}\psi_{ABCD}$ differs from $\psi_{ABCDE\cdots K}{}^{P'\cdots V'}$ by expressions obtained by applying rule (5.5) and (5.6) to ψ_{ABCD} and derivatives of ψ_{ABCD}, and perhaps differentiating further. Since $\lambda = 0$, this leads to terms of the form

$$\psi_{XAG\cdots K}{}^{T'\cdots V'}\psi^{X}{}_{B\cdots F}{}^{P'\cdots S'} \quad \text{or} \quad \bar{\psi}_{Y'}{}^{P'T'\cdots V'}{}_{G\cdots K}\psi_{A\cdots F}{}^{YQ'\cdots S'}$$

only. (By the inductive hypothesis all the derivatives of ψ_{ABCD} which occur are equal to the $\psi\ldots{}^{\cdots}$'s.) These terms all involve contractions between the $\psi\ldots{}^{\cdots}$'s. But with $\psi\ldots{}^{\cdots}$'s given by (5.9), any contraction must clearly vanish (since $\pi_X\pi^X = 0$). Hence

$$(\partial_E{}^{P'} \cdots \partial_K{}^{V'}\psi_{ABCD})_0 = \alpha_r\pi_A\pi_B \cdots \pi_K\bar{\pi}^{P'} \cdots \bar{\pi}^{V'}$$

as required.

The curvature at points other than O can now be calculated:

$$(\psi_{ABCD})_x = \alpha_0\pi_A\pi_B\pi_C\pi_D + \alpha_1 x^{EP'}\pi_A \cdots \pi_E\bar{\pi}_{P'} + \frac{1}{2!}\alpha_2 x^{EP'}x^{FQ'}\pi_A \cdots \pi_F\bar{\pi}_{P'}\bar{\pi}_{Q'} + \cdots$$

$$= f(x^{EP'}\pi_E\bar{\pi}_{P'})\pi_A\pi_B\pi_C\pi_D = f(x^\mu p_\mu)\pi_A\pi_B\pi_C\pi_D,$$

where

$$f(s) = \alpha_0 + \alpha_1 s + \frac{1}{2!}\alpha_2 s^2 + \frac{1}{3!}\alpha_3 s^3 + \cdots \tag{5.10}$$

and $p_{AB'} = \pi_A\bar{\pi}_{B'}$. Thus the curvature is a function of the one parameter $x^\mu p_\mu$ only. It is constant along the (null) 3-spaces $x^\mu p_\mu =$ constant. Furthermore, by (5.2),

$$(\partial_{EP'}\psi_{ABCD})_x = (\partial_{EP'}\psi_{ABCD})_0 + x^{FQ'}(\partial_{FQ'}\{\partial_{EP'}\psi_{ABCD}\})_0 + \frac{1}{2!}x^{FQ'}x^{GR'}(\partial_{FQ'}\partial_{GR'}\{\partial_{EP'}\psi_{ABCD}\})_0 + \cdots$$

$$= f'(x^\mu p_\mu)\pi_A\pi_B\pi_C\pi_D\pi_E\bar{\pi}_{P'},$$

$$(\partial_{EP'}\partial_{FQ'}\psi_{ABCD})_x = f''(x^\mu p_\mu)\pi_A \cdots \pi_F\bar{\pi}_{P'}\bar{\pi}_{Q'},$$

etc. Hence ψ_{ABCD}, $\psi_{ABCDEP'}$, $\psi_{ABCDEFP'Q'}$, $\cdots$ are all constant along the 3-space $x^\mu p_\mu = 0$. It follows that the whole space $\mathfrak{M}$ admits the three-parameter group of

translations[16] in the directions lying in this 3-space. The space $\mathfrak{M}$ thus represents a plane wave which moves uniformly with the velocity of light in the direction represented by p_μ. The intensity and polarization of the wave are determined by the modulus and argument of the function $f(s)$.

Particular cases of interest are:

(i) the constant gravitational field with ψ_{ABCD} constant everywhere. Here $f(s)$ = constant, i.e., $\alpha_1 = \alpha_2 = \cdots = 0$, and $\mathfrak{M}$ admits additional translational motions.

(ii) Sinusoidal waves;

$$f(s) \equiv ae^{ins} + be^{ins}, \quad \text{i.e.,} \quad \alpha_r = a(in)^r + b(-in)^r.$$

In this case $\mathfrak{M}$ admits an additional discrete group of translations.

(iii) Gravitational pulse; for example,

$$f(s) = \begin{cases} b \exp\left(\dfrac{c}{s-a} - \dfrac{c}{s+a}\right) & \text{if } -a < s < a \\ 0 & \text{if } s \leqq -a \text{ or } s \geqq a. \end{cases}$$

Case (iii) is not strictly an analytic manifold. $\mathfrak{M}$ has to be constructed from three analytic pieces (two of which are flat). The middle piece fits on smoothly to the other two pieces, the join being C^∞. The space is exactly flat before the pulse arrives and is again exactly flat after the pulse has departed (*23*, p. 523).

An advantage of a method such as this for obtaining spaces satisfying Einstein's equations is that the usual problem of deciding whether an effect is real or merely due to a bad choice of coordinates simply does not arise. The curvature at any point is found directly. However, it will naturally be convenient to be able to introduce coordinates into a space defined in this way, if desired. A coordinate system on $\mathfrak{M}$ may be thought of as a set of four scalar fields $u_{(i)}$ $(i = 0, \cdots 3)$. The symmetric derivatives $\partial_{(\alpha} \cdots \partial_{\gamma)} u_{(i)}$ of each $u_{(i)}$ may be specified arbitrarily at the point 0. The values of the coordinates $u_{(i)}$ and their derivatives at any other point may then be calculated using (5.2), after some of the unsymmetrized derivatives have been obtained using (2.1). The expression for the metric at each point can be obtained from the first derivatives of the $u_{(i)}$ at that point. This method will be described in detail in a later paper.

The case when an electromagnetic field is present in the space can be treated by an extension of the coordinate-free method for empty space described above. The spinors

$$\psi_{ABCDE\cdots G}{}^{P'\cdots R'} = \partial_{(E}{}^{(P'} \cdots {}_G{}^{R')} \psi_{ABCD)}$$

[16] $\mathfrak{M}$ also admits a two parameter group of rotational (Lorentz) symmetries given by the unimodular matrices $t^A{}_B$ satisfying $t^A{}_B \pi^B = \pm \pi^A$, and disconnected from these, the rotations for which $t^A{}_B \pi^B = \pm i\pi^A$. There may also be some reflectional symmetries in special cases. This five parameter group of motions serves to characterize the plane wave solutions (see Bondi *et al.*, (*23*)).

are defined as before and spinors ϕ_{AB} , $\phi_{ABC}{}^{P'}$, $\phi_{ABCD}{}^{P'Q'}$, $\cdots$ are introduced, defined similarly by

$$\phi_{ABC\cdots E}{}^{P'\cdots R'} = \partial_{(C}{}^{(P'} \cdots \partial_{E}{}^{R')}\phi_{AB)} .$$

By the same kind of argument as before, it follows that ϕ_{AB} , $\phi_{ABC}{}^{P'}$, $\cdots$, ψ_{ABCD} , $\psi_{ABCDE}{}^{P'}$, $\cdots$ are all algebraically independent. Instead of (5.4) we have

$$\epsilon^{CA}\partial_{C}{}^{P'}\phi_{AB} = 0 \quad \text{and} \quad -\epsilon^{EA}\partial_{E}{}^{P'}\psi_{ABCD} = \bar{\phi}^{P'}{}_{Q'}\partial_{D}{}^{Q'}\phi_{BC}$$

from (3.10) and (3.12). The first of these states the symmetry of

$$\partial_{C}{}^{P'} \cdots \partial_{E}{}^{R'}\phi_{AB}$$

in E, A, while the second expresses the part of

$$\partial_{E}{}^{P'} \cdots \partial_{G}{}^{R'}\psi_{ABCD}$$

skew in G, A in terms of derivatives of ϕ_{AB} of at most the same order. They imply no condition on the symmetrized derivatives of ϕ_{AB} or ψ_{ABCD} . Nor do the equivalents of (5.5) and (5.6), which differ from them only in that the second relation (5.5) is replaced by

$$\epsilon_{R'S'}\{\partial_{G}{}^{R'}\partial_{H}{}^{S'} + \partial_{H}{}^{R'}\partial_{G}{}^{S'}\}\eta^{P'} = \phi_{GH}\bar{\phi}^{P'}{}_{Q'}\eta^{Q'}$$

(see 3.13) and the first relation (5.6) by

$$\epsilon^{GH}\{\partial_{G}{}^{R'}\partial_{H}{}^{S'} + \partial_{G}{}^{S'}\partial_{H}{}^{R'}\}\xi_{A} = \bar{\phi}^{R'S'}\phi_{AB}\xi^{B}$$

The argument to show that the unsymmetrized derivatives can be expressed algebraically in terms of the symmetrized derivatives is exactly analogous to that for pure gravitational case. The derivative $\partial_{C}{}^{P'} \cdots \partial_{E}{}^{R'}\phi_{AB}$ differs from $\phi_{ABC\cdots E}{}^{P'\cdots R'}$ by expressions constructed from lower order derivatives of ϕ_{AB} and ψ_{ABCD} , while $\partial_{E}{}^{P'} \cdots \partial_{G}{}^{R'}\psi_{ABCD}$ differs from $\psi_{ABCDE\cdots G}{}^{P'\cdots R'}$ by expressions constructed from derivatives of ϕ_{AB} of the same order or lower and from lower order derivatives of ψ_{ABCD} . Thus, we can construct $\partial_{C}{}^{P'}\phi_{AB}$, $\partial_{E}{}^{P'}\psi_{ABCD}$, $\partial_{C}{}^{P'}\partial_{D}{}^{Q'}\phi_{AB}$, $\partial_{E}{}^{P'}\partial_{F}{}^{Q'}\psi_{ABCD}$, $\cdots$, in that order, from the symmetrized derivatives. The symmetric spinors ϕ_{AB} , $\phi_{ABC}{}^{P'}$, $\phi_{ABCD}{}^{P'Q'}$, $\cdots$, ψ_{ABCD} , $\psi_{ABCDE}{}^{P'}$, $\cdots$ can therefore be specified arbitrarily at a point 0 (apart from convergence considerations) and ϕ_{AB} , ψ_{ABCD} at any other point can be determined from them by (5.2).

A simple example is the case of a combined gravitational-electromagnetic wave (see also *22*). Here ψ_{ABCD} , $\psi_{ABCDEP'}$, $\cdots$ are given by (5.9) and

$$\phi_{AB} = \beta_0\pi_A\pi_B , \quad \phi_{ABCP'} = \beta_1\pi_A\pi_B\pi_C\bar{\pi}_{P'} ,$$

$$\phi_{ABCDP'Q'} = \beta_2\pi_A\pi_B\pi_C\pi_D\bar{\pi}_{P'}\bar{\pi}_{Q'} , \cdots$$

at the point O. As was the case, considered earlier, with the pure gravitational

TABLE I

SUMMARY OF SOME OF THE RESULTS OF THIS PAPER ON THE COMPARISONS BETWEEN ELECTROMAGNETIC AND GRAVITATIONAL FIELDS IN SPINOR FORM

	Maxwell field	Curvature tensor with $R^{\mu}{}_{\gamma\mu\sigma}=0$
Tensor-spinor correspondence	$F_{\mu\gamma} \leftrightarrow \frac{1}{2}\{\phi_{AB}\epsilon_{C'D'} + \epsilon_{AB}\bar{\phi}_{C'D'}\}$	$R_{\mu\gamma\rho\sigma} \leftrightarrow \frac{1}{2}\{\psi_{ABCD}\epsilon_{E'F'}\epsilon_{G'H'} + \epsilon_{AB}\epsilon_{CD}\bar{\psi}_{E'F'G'H'}\}$
First order equation	Maxwell equations: $\partial^{AC'}\phi_{AB} = 0$	Bianchi identities: $\partial^{AE'}\psi_{ABCD} = 0$
(Super-)energy tensor	Maxwell stress tensor $\leftrightarrow$ $\frac{1}{2}\phi_{AB}\bar{\phi}_{C'D'}$	Robinson-Bel tensor $\leftrightarrow$ $\psi_{ABCD}\bar{\psi}_{E'F'G'H'}$
Duality rotations	$\phi_{AB} \to e^{i\theta}\phi_{AB}$	$\psi_{ABCD} \to e^{i\theta}\psi_{ABCD}$
Canonical representation	$\phi_{AB} = \eta_{(A}\zeta_{B)}$	$\psi_{ABCD} = \alpha_{(A}\beta_B\gamma_C\delta_{D)}$
Classification scheme	[11] $K \neq 0$ [2] $\to$ [—] $K = 0$	[1111] $I^3 \neq 6J^2$ [211] $\to$ [22] $I^3 = 6J^2 \neq 0$ [31] $\to$ [4] $\to$ [—] $I = J = 0$
Plane wave	$\phi_{AB}(x^{\mu}) = g(x^{\mu}p_{\mu})\pi_A\pi_B$	$\psi_{ABCD}(x^{\mu}) = f(x^{\mu}p_{\mu})\pi_A\pi_B\pi_C\pi_D$

wave, the unsymmetrized derivatives of ϕ_{AB} and ψ_{ABCD} turn out to be equal to the symmetrized derivatives provided that $\lambda = 0$. Hence

$$(\phi_{AB})_x = g(x^{\mu}p_{\mu})\pi_A\pi_B\,, \quad (\psi_{ABCD})_x = f(x^{\mu}p_{\mu})\pi_A\pi_B\pi_C\pi_D\,,$$

where

$$g(s) \equiv \beta_0 + \beta_1 s + \frac{1}{2!}\beta_2 s^2 + \cdots$$

and $f(s)$ is given by (5.10) as before. The discussion given in the pure gravitational case applies here also. The function $g(s)$ determines the intensity and polarization of the electromagnetic part of the wave and $f(s)$ the "purely gravitational" part. The electromagnetic field is null everywhere and the gravitational field is [4]. All six principal null directions coincide and point in the direction p_{μ} giving the motion of the wave.

Table I summarizes some of the many analogies between the electromagnetic and gravitational fields, that are brought out by the spinor formalism.

I should like to offer my thanks to Dr. D. W. Sciama for his early encouragement and for many invaluable discussions.

RECEIVED: September 16, 1959

REFERENCES

1. L. Infeld and B. L. van der Waerden, *Sitzber. preuss. Akad. Wiss. Physik.-math. Kl.* **9:** 380 (1933).
2. W. L. Bade and H. Jehle, *Revs. Modern Phys.* **25:** 714 (1953).
3. H. S. Ruse, *Proc. Roy. Soc. Edinburgh* **57:** 97 (1937).
4. G. Y. Rainich, *Trans. Am. Math. Soc.* **27:** 106 (1925).
5. C. W. Misner and J. A. Wheeler, *Annals of Physics* **2:** 525 (1957).
6. L. Witten, *Phys. Rev.* **113:** 357 (1959).
7. P. A. M. Dirac, *Proc. Roy. Soc.* **A155:** 447 (1936).
8. L. Bel, *Compt. rend.* **247:** 1094 (1958).
9. L. Bel, *Compt. rend.* **248:** 1297 (1959).
10. E. M. Corson, "Introduction to Tensors, Spinors, and Relativistic Wave-equations." Blackie, London, 1953.
11. A. S. Eddington, "The Mathematical Theory of Relativity." Cambridge Univ. Press, London and New York, 1923.
12. J. L. Synge, "Relativity: The Special Theory." North-Holland Publ. Co., Amsterdam, 1956.
13. L. Witten, *Phys. Rev.* **115:** 206 (1959).
14. H. S. Ruse, *Proc. Roy. Soc. Edinburgh* **A62:** 64 (1944).
15. R. Debever, *Compt. Rend. Acad. Sci.* **249:** 1324 (1959).
16. R. Debever, *Compt. Rend. Acad. Sci.* **249:** 1744 (1959).
17. F. A. E. Pirani, *Phys. Rev.* **105:** 1089 (1957).
18. J. H. Grace and A. Young, "Algebra of Invariants." Cambridge Univ. Press, London and New York, 1903.
19. R. Penrose, *Proc. Cambridge Phil. Soc.* **55:** 137 (1959).
20. A. Z. Petrov, *Doklady Akad. Nauk, SSSR* **105:** 905 (1955).
21. J. Géhéniau, *Compt. rend.* **244:** 723 (1957).
22. A. Lichnerowicz, *Compt. rend.* **246:** 893 (1958).
23. H. Bondi, F. A. E. Pirani and I. Robinson, *Proc. Roy. Soc.* **A251:** 519 (1959).
24. F. A. E. Pirani, *Proc. Roy. Soc.* **A252:** 96 (1959).
25. T. Y. Thomas, "Differential Invariants of Generalized Spaces." Cambridge Univ. Press, London and New York, 1934.

13
IMBEDDING OF MANIFOLDS IN EUCLIDEAN SPACE

Annals of Math. (2) **73** (1961), 613–623. With J. H. C. Whitehead and E. C. Zeeman

Apparently, the basic procedure put forward here became known as the 'engulfing' method.

ANNALS OF MATHEMATICS
Vol. 73, No. 3, May, 1961
Printed in Japan

IMBEDDING OF MANIFOLDS IN EUCLIDEAN SPACE

BY R. PENROSE, J. H. C. WHITEHEAD*, AND E. C. ZEEMAN

(Received June 17, 1960)

1. The main theorems

It is obvious, since the vertices may be placed in general position, that a finite, n-dimensional simplicial complex can be piecewise linearly imbedded in euclidean $(2n + 1)$-space, R^{2n+1}. This is the best possible result for an arbitrary complex since the n-skeleton of a $(2n + 2)$-simplex cannot be imbedded in R^{2n} [4], [9]. On the other hand a compact, smooth or combinatorial n-manifold (see § 2 for definitions) can be (smoothly or piecewise linearly) imbedded in R^{2n} [13], [19], [20]. Real projective n-space cannot be smoothly imbedded in R^{2n-1} if $n = 2^k$ [14], [21], though there are better results for certain other projective spaces [8].

In this paper we are concerned with piecewise linear imbeddings in R^q of compact, n-dimensional, combinatorial manifolds (see § 2) which are $(m - 1)$-connected, where $0 < 2m \leqq n$. The condition $m > 0$ means that such a manifold is connected. If a closed (i.e., compact, unbounded) n-manifold M is $(m - 1)$-connected and $2m > n$, then it follows from the Poincaré duality that M has the homotopy type of an n-sphere. Therefore, if it turns out that every such manifold is a (combinatorial) n-sphere, or even if it can be piecewise linearly imbedded in R^{n+1}, then (1.1) below is valid for $0 < m \leqq n$. Except when the contrary is stated, it is to be understood that all the manifolds to which we refer are combinatorial and that all our maps, in particular the immersions (see § 2) and imbeddings, are piecewise linear. We prove:

THEOREM (1.1). *If $0 < 2m \leqq n$, then every closed, $(m - 1)$-connected n-manifold can be imbedded in R^{2n-m+1}.*

THEOREM (1.2). *Let M be a compact, bounded n-manifold which is $(m - 1)$-connected $(0 < 2m \leqq n)$. If either*

(a) *$\dot{M} \times I$ can be imbedded in R^{2n-m}, or*

(b) *$\dot{M}$ is $(m - 2)$-connected ((-1)-connected means non-vacuous), then M can be imbedded in R^{2n-m}.*

It follows from (2.3) that the condition (a) is necessary for the imbeddability of M in R^{2n-m}. It is obviously satisfied if $\dot{M}$ can be imbedded in R^{2n-m-1}, hence, by (1.1), if each component of $\dot{M}$ is $(m - 1)$-connected and $2m < n$.

* Deceased May 8, 1960.

By the *branch locus* of a map $f\colon M \to R^q$ we mean the set of points $x \in M$ such that no neighbourhood of x is imbedded by f. Let K be a (rectilinear) triangulation of M such that f is barycentric in each simplex of K (we do not assume that f is simplicial with respect to K and a triangulation of R^q). Then the branch locus, B, of f is the union of all the closed simplexes $\sigma \in K$ such that $f \mid \mathrm{St}(\sigma, K)$ is not an imbedding, where $\mathrm{St}(\sigma, K)$ denotes the union of all the closed simplexes of K which contain σ. Therefore B is a compact polyhedron. Clearly $f \mid M - B$ is an immersion. We shall prove:

THEOREM (1.3). *Let M be a closed, $(m-1)$-connected n-manifold, where $0 < 2m \leqq n$. Assume that there is a map $M \to R^q$ whose branch locus is at most $(m-1)$-dimensional, where $q > 2(n-m)$. Then M can be imbedded in R^{q+1}.*

In particular M can be imbedded in R^{q+1} if it can be immersed in R^q, provided M satisfies the conditions of (1.3) and $q > 2(n-m)$. Thus if $n = 2m$ and M can be immersed in R^q (whence $q > n$ because M is closed), then it can be imbedded in R^{q+1}.

PROOF OF (1.1), *assuming* (1.3). Let $f\colon M \to R^{2n-m}$ be a map which is barycentric in each simplex of a triangulation K of M and which maps the vertices of K in general position. Then f imbeds each simplex of K. Let σ_1, σ_2 be simplexes of K, let $\sigma_0 = \sigma_1 \cap \sigma_2$ and let $\dim \sigma_0 = p \geqq m$, $\dim \sigma_1 = r$, $\dim \sigma_2 = s$. Then σ_1 is the join of σ_0 and an $(r-p-1)$-simplex τ. Since

$$(r-p-1) + s - (2n-m) \leqq m - p - 1 < 0$$

and the vertices of K are mapped in general position, it follows that f does not meet the s-plane containing $f\sigma_2$. Therefore $f\sigma_1 \cap f\sigma_2 = f\sigma_0$ and it follows that $f \mid \mathrm{St}(\sigma, K)$ is an imbedding if $\dim \sigma \geqq m$ $(\sigma \in K)$. Therefore the branch locus of f is at most (m − 1)-dimensional and (1.1) follows from (1.3) since $2n - m > 2(n-m)$.

Let P, Q be compact polyhedra in a manifold M. We describe Q as *quasi-complementary* to P if, and only if, every compact polyhedron in $M - P$ can be (piecewise linearly) imbedded in every neighbourhood of Q. For example, let K be a triangulation of M and let the vertices of K be separated into two disjoint subsets A, B. Let P be the union of the simplexes of K whose vertices are all in A and Q the union of the simplexes of K whose vertices are all in B. Then P, Q are quasi-complementary to each other [2]. In particular, if L is the cell-complex dual to K, then it follows from the preceding remark, applied to the first barycentric subdivision of K, that $|L^{n-m}|$ is quasi-complementary to $|K^{m-1}|$

(X^r denotes the r-skeleton of a given complex X and $|X|$ denotes the polyhedron covered by a given complex X). In general P and Q may have points in common. For example, if M is a (combinatorial) n-sphere, then P, Q may be any proper, non-vacuous, compact, polyhedral subsets of M.

THEOREM (1.4). *Let M be a compact (possibly bounded), $(m-1)$-connected, n-manifold, where $0 < 2m \leqq n$. Assume that there are compact polyhedra $P \subset \operatorname{int} M$, $Q \subset M$ such that $\dim P < m$, Q is quasi-complementary to P and some neighbourhood of Q can be imbedded in R^q. Then M can be imbedded in R^{q+1} and in R^q if it is bounded.*

PROOF. Let $U \subset M$ be a neighbourhood of Q which can be imbedded in R^q. Since $P \subset \operatorname{int} M$, $\dim P < m$ and M is $(m-1)$-connected it follows from (2.9) below that, if M is bounded, it can be imbedded in $M - P$, hence in U and hence in R^q.

Let M be unbounded. Then, by (2.7), there is an n-element $E \subset M$ such that $P \subset \operatorname{int} E$. Let $M_0 = M - \operatorname{int} E$. Then M_0 can be imbedded in U and there is therefore an imbedding $f\colon M_0 \to R^q$. We take R^q to be a hyperplane in R^{q+1} and extend f to an imbedding $M \to R^{q+1}$, which maps E on the join of $f\dot{E}$ and a point in $R^{q+1} - R^q$. This proves (1.4).

LEMMA (1.5). *Let Q be a compact polyhedron in a manifold M and let Q have a neighbourhood which can be immersed in R^q, where $q > 2 \dim Q$. Then Q has a neighbourhood which can be imbedded in R^q.*

This follows from the properties of general position and from (2.1).

PROOF OF (1.3). Let $f\colon M \to R^q$ be a map whose branch locus, B, is at most $(m-1)$-dimensional and let K be a triangulation of M such that f is barycentric in each simplex of K. Let L be the cell-complex on M which is dual to K. Then $f|M - B$ is an immersion, $B \subset K^{m-1}$ and (1.3) follows from (1.5), (1.4) with $P = |K^{m-1}|$, $Q = |L^{n-m}|$.

Let M be a homotopy n-sphere (i.e., a combinatorial manifold of the homotopy type of an n-sphere), let E be an n-element in M and let $M_0 = M - \operatorname{int} E$. Then it follows from a theorem in a forthcoming paper by A. M. Gleason that there is an immersion $f\colon M_0 \to R^n$. Let $h\colon E \to \Delta$ be a homeomorphism of E on an n-simplex Δ, let $a \in \operatorname{int} \Delta$, $b \in R^{n+1} - R^n$ and let $k\colon \Delta \to R^{n+1}$ be defined by

$$k((1-t)a + thx) = (1-t)b + tfx \qquad (x \in \dot{E},\ t \in I)\ .$$

Then a map $g\colon M \to R^{n+1}$ is defined by $gx = fx$ or khx according as $x \in M_0$ or $x \in E$. Since f is an immersion it follows that the branch locus of g consists, at most, of the point $h^{-1}a$. Therefore we have, by (1.3):

THEOREM (1.6). *A homotopy n-sphere can be imbedded in R^{n+2} if $n = 2m$ and in R^{n+3} if $n = 2m+1$.*

We describe M as *combinatorially equivalent* to a smooth manifold M_1 if, and only if, some triangulation, K, of M is the argument complex in a C^1-triangulation of M_1 [17]. This means that there is a homeomorphism $f\colon M \to M_1$ such that, for every closed n-simplex $\sigma \in K$, the map $f \mid \sigma$ can be extended to a regular C^1-imbedding $U(\sigma) \to M_1$, where $U(\sigma)$ is an open neighbourhood of σ in the n-plane which contains it. The manifold M_1 is said to be *almost parallelizable* if, and only if, $M_1 - p$ is parallelizable for some, and therefore every point $p \in M_1$. Let M be closed and let K and $f\colon M \to M_1$ be as above, where M_1 is almost parallelizable. Let M be $(m-1)$-connected $(0 < 2m \leqq n)$, let $E \subset M$ be an n-element such that $K^{m-1} \subset \operatorname{int} E$ and let $M_0 = M - \operatorname{int} E$. Then fM_0 is parallelizable. Therefore there is a smooth immersion $fM_0 \to R^{n+1}$ [6, p. 269]. Let K_0 be a triangulation of M_0 such that, for every simplex $\sigma \in K_0$, the restriction of $fM_0 \to R^{n+1}$ to some neighbourhood of $f\operatorname{St}(\sigma, K_0)$ is a regular imbedding. Then it follows from Theorems 1 and 3 in [17] (Theorem 3 may be applied to every $\operatorname{St}(\sigma, K_0)$) that M_0 can be piecewise linearly immersed in R^{n+1} and hence in $R^{2(n-m)+1}$. Let L be the cell-complex dual to K and let $Q = |L^{n-m}|$. Then, as in the proof of (1.3), M_0 can be imbedded in every neighbourhood of Q and some neighbourhood of Q can be imbedded in $R^{2(n-m)+1}$. Therefore we have, by an argument in the proof of (1.4):

THEOREM (1.7). *Let M be a closed, $(m-1)$-connected n-manifold $(0 < 2m \leqq n)$, which is combinatorially equivalent to an almost parallelizable smooth manifold. Then M can be piecewise linearly imbedded in $R^{2(n-m)+2}$.*

Theorems (1.6), (1.7) were pointed out to us by M. W. Hirsch.

PROOF OF (1.2). Let $\dot{M} \times I$ be imbeddable in R^{2n-m}. By (2.3) there is a closed, polyhedral neighbourhood $N \subset M$ of $\dot{M}$ which is (piecewise linearly) homeomorphic to $\dot{M} \times I$ and hence imbeddable in R^{2n-m}. Let K be a triangulation of the pair (M, N) (i.e., a triangulation of M with a subcomplex covering N) such that a given imbedding $N \to R^{2n-m}$ is barycentric in each simplex of $K \cap N$. Extend this to a map $f\colon K \to R^{2n-m}$ which is barycentric in each simplex of K. We assume, as we obviously may, that $\dot{K}$ is a *full subcomplex* of K (i.e., that every simplex of K with all its vertices in $\dot{K}$ is contained in $\dot{K}$) and that f maps the vertices of K in general position. Let P denote the union of all the closed simplexes of K which are at most $(m-1)$-dimensional and do not meet $\dot{M}$. Then the branch locus of f is contained in P.

Let K' denote the first barycentric subdivision of K. The simplexes of K' are of the form $c(\sigma_0)\cdots c(\sigma_p)$, where $c(\sigma)$ denotes the centroid of a given simplex $\sigma \in K$ and $\sigma_i \subset \sigma_{i+1}$ $(i = 0, \cdots, p-1)$. Let $K_0 \subset K'$ be the

subcomplex which consists of all simplexes $c(\sigma_0)\cdots c(\sigma_p) \in K'$ such that $\sigma_0 \not\subset \dot{K}$. Then $K_0 = K' - \mathrm{O}(\dot{K}', K')$, where $\mathrm{O}(\dot{K}', K')$ denotes the union of the open simplexes of K' whose closures meet $\dot{K}'$. Let $M_0 = |K_0|$. Since $\dot{K}$ is a full subcomplex of K it follows from (2.3), applied to both M and M_0, that M_0 is homeomorphic to M. Therefore it is enough to prove that M_0 can be imbedded in R^{2n-m}.

Let us describe a vertex $c(\sigma) \in K_0$ as of the *first kind* if $\dim(\sigma) < m$ and $\sigma \cap \dot{K} = \emptyset$ (σ denotes a closed simplex) and of the *second kind* if either $\dim(\sigma) \geqq m$ or $\sigma \cap \dot{K} \neq \emptyset$. Then the polyhedron P is the union of the simplexes of K_0 whose vertices are all of the first kind. Let Q denote the union of the simplexes of K_0 whose vertices are all of the second kind. Then Q is quasi-complementary to P. It consists of all simplexes $c(\sigma_0)\cdots c(\sigma_p) \in K_0$ such that either $\dim(\sigma_0) \geqq m$, whence $p \leqq n - m$, or $\sigma_0 \cap \dot{K} \neq \emptyset$, which means that $c(\sigma_0)\cdots c(\sigma_p) \in \dot{K}_0$. Moreover $Q \subset M_0 - P$ and $\dot{M}_0 \subset N$, whence $f \mid \dot{M}_0$ is an imbedding.

Let the images of the vertices of K_0 be shifted slightly so as to define a map $f_0\colon M_0 \to R^{2n-m}$, barycentric in each simplex of K_0, such that $f_0 \mid \dot{M}_0$ is an imbedding, $f_0 \mid M - P$ is an immersion and f_0 maps the vertices of K_0 in general position. Since $\dim(Q - \dot{M}_0) \leqq n - m$ and $(n-m)+(n-1) < 2n - m$ it follows the $f_0 \mid Q$ is an imbedding. By (2.1) f_0 imbeds some neighbourhood of Q and (1.2) with Hypothesis (a) follows from (1.4).

Now let $\dot{M}$ be $(m-2)$-connected and let M_1 be a copy of M such that $\dot{M}_1 = \dot{M} = M \cap M_1$. Let $M_2 = M \cup M_1$. Then it follows, trivially if $m = 1$ (since $\dot{M} \neq \emptyset$) and from a theorem due to van Kampen [9, p. 177] and from the Mayer-Vietoris theorem [3] if $m > 1$, that M_2 is $(m-1)$-connected. Let E be an n-element in M_1 and let $M_3 = M_2 - \operatorname{int} E$. Then M_3 is $(m-1)$-connected and $\dot{M}_3$ is the $(n-1)$-sphere $\dot{E}$. Therefore it follows from (1.2) with Hypothesis (a) that M_3 can be imbedded in R^{2n-m}. Since $M \subset M_3$ this completes the proof.

The problem of adapting these methods, for closed manifolds, to the smooth theory leads to the following question. Let D^n be the n-disc bounded by a unit $(n-1)$-sphere $S^{n-1} \subset R^n$ and let $f\colon S^{n-1} \to R^q$ be a regular imbedding of class C^r $(1 \leqq r < \infty)$. Can f be extended to a regular C^r imbedding $g\colon D^n \to R^{q+1}$ such that $g(\operatorname{int} D^n) \subset R^{q+1} - R^q$? The answer is "yes" if $q \geqq 2n$, but there is an unpublished theorem due to R. H. Fox and J. W. Milnor (see [5]) which implies that, if $n = 2$, $q = 3$, there are imbeddings f for which the answer is "no". This is because fS^{n-1} is knotted in a certain way and does not bound any regular C^r disc whose interior lies in $R^{q+1} - R^q$. For larger values of n and $q < 2n$ there may, possibly, be cases in which the answer is

"no" even though fS^{n-1} is "good", say $fS^{n-1} = S^{n-1}$ $(R^n \subset R^q)$, because f is "bad". (Cf. [10].) For the case of imbeddings of bounded manifolds these difficulties are not always so serious. Some results along these lines have recently been obtained by M. W. Hirsch [7].

Many of the results given here lead to imbeddings which are *locally unknotted*, however [22]. (The piecewise linear imbedding $f\colon M \to R^q$ is locally unknotted if, for triangulations K of fM and L of R^q with K a subcomplex of L, we have $\dot{\mathrm{St}}\,(v, K)$ unknotted in $\dot{\mathrm{St}}\,(v, L)$ for every vertex $v \in K$. That is, $\dot{\mathrm{St}}\,(v, K) = \dot{E}$ for some n-element $E \subset \dot{\mathrm{St}}\,(v, L)$.) It follows from the results of [22] that any (piecewise linear) imbedding of an n-manifold M in R^q must be locally unknotted if $2(q-1) > 3n$. Thus, in particular, the imbedding of (1.1) is necessarily locally unknotted unless $2m = n$. Also, it is not hard to adapt the proof of (1.2) to obtain locally unknotted imbeddings in all its cases (although the case $m = 1$, $n = 3$ appears to need special consideration).

2. Definitions and lemmas

A map $f\colon X \to Y$, where X, Y are arbitrary topological spaces, is called an *imbedding* if, and only if, it is a homeomorphism onto fX. A map $f\colon X \to Y$ is called an *immersion* if, and only if, every point $x \in X$ has a neighbourhood $N_x \subset x$ such that $f \mid N_x$ is an imbedding.

LEMMA (2.1). *Let $f\colon X \to Y$ be an immersion of a locally compact metric space X in a Hausdorff space Y and let $f \mid A$ be an imbedding, where A is a compact subset of X. Then there is a compact neighbourhood $N \subset X$ of A such that $f \mid N$ is an imbedding.*

PROOF. Let $N_i = \{x \in X \mid \delta(x, A) \leqq 1/i\}$ $(i = 1, 2, \cdots)$, where δ is a metric for X. Then there is an integer k such that N_i is compact if $i \geqq k$. Assume that $f \mid N_i$ is not an imbedding for any i. Then, there are points $x_i, x_i' \in N_i$ such that $x_i \neq x_i'$, $fx_i = fx_i' = y_i$, say, because N_i is compact if $i \geqq k$ and Y is a Hausdorff space. Since f is locally 1-1 and $f \mid A$ is 1-1 it follows without difficulty that some subsequence of the sequence $\{y_i\}$ converges to each of two distinct points in fA. This is absurd and (2.1) follows.

By a (compact) polyhedron we mean a subspace of R^q, for some q, which can be triangulated by a finite, rectilinear, simplicial complex. It is to be understood that all the triangulations of polyhedra and subdivisions of complexes to which we refer are rectilinear. A map $P \to Q$, where P, Q are polyhedra, is called *piecewise linear* if, and only if, it is simplical with respect to suitable triangulations of P, Q. Thus P, Q are piecewise linearly homeomorphic if, and only if, they have isomorphic triangulations.

This is an equivalence relation because two triangulations of the same polyhedron have a common subdivision. More generally, if P_0, P are polyhedra such that $P_0 \subset P$ and if K_0, K are triangulations of P_0, P, then there is a subdivision of K with a subcomplex which is a subdivision of K_0 [15]. As stated in § 1, all the maps between polyhedra to which we refer will be piecewise linear. Thus "homeomorphic", with reference to polyhedra, will always mean "piecewise linearly homeomorphic". If K is a triangulation of a polyhedron P and if X denotes either a subset of P or a subcomplex of K, then $N(X, K)$ will denote the union of all the closed simplexes of K which meet X. The symbol $N(X, K)$ will denote either a polyhedron or a complex, according to the context (or the choice of the reader).

By an *n-element* (*n-sphere*) we mean a polyhedron which is homeomorphic to a closed n-simplex (boundary of an $(n+1)$-simplex). By a (*combinatorial*) *n-manifold* we mean a polyhedron, M, with a triangulation K such that St (v, K) is an n-element, for every vertex $v \in K$. This property is independent of the choice of K. We denote the boundary of a manifold M by $\dot{M}$ and int $M = M - \dot{M}$.

Let M be a bounded n-manifold and let $E^n \subset M$ be an n-element such that $\dot{M} \cap E^n = E^{n-1}$, say, is an $(n-1)$-element in $\dot{E}^n$. Then $\dot{E}^n - \text{int } E^{n-1}$ is also an $(n-1)$-element [1, Theorem 14.2]. Let $M_0 = M - \text{int } E^n - \text{int } E^{n-1}$. Then we have [11, Theorem 8a], [1, Theorem (14.3)]:

LEMMA (2.2). *M_0 is homeomorphic to M.*

Let M be as in (2.2) and let K be a triangulation of M such that $\dot{K}$ is a full subcomplex of K (see the proof of (1.2)) and let K' denote the first barycentric subdivision of K. Then we have [16]:

THEOREM (2.3). *$N(\dot{M}, K')$ is homeomorphic to $\dot{M} \times I$.*

Let A, B be polyhedra such that $B = A \cup E$, where E is a k-element $(k > 0)$ and $A \cap E$ is a $(k-1)$-element in $\dot{E}$. Then the ordered pair (A, B) will be called an *elementary expansion* (of order k) and (B, A) an *elementary contraction* (of order k). A polyhedron P will be said to *expand* into Q, and Q to *collapse* into P, if, and only if, either $P = Q$ or there is a sequence of elementary expansions (A_i, A_{i+1}) $(i = 1, \cdots, r-1)$, of arbitrary orders, such that $A_1 = P$, $A_r = Q$. A polyhedron P will be called *completely collapsible* if, and only if, it collapses to a point. Obviously an element is completely collapsible. Let P be a polyhedron in an n-manifold M. By a *regular enlargement* (in M) of P we mean an n-manifold, N, such that $P \subset N \subset M$ and N collapses into P. By a *regular neighbourhood* of P we mean a regular enlargement $N \subset M$ of P which is a closed

neighbourhood of P (i.e., $P \cap \overline{M - N} = \varnothing$). If P is an n-manifold it is a regular enlargement of itself. Therefore (2.2) is a special case of (2.5) below (the proof of (2.5) depends on (2.2)).

Let K be a triangulation of the pair (M, P) and let K'' be its second barycentric subdivision. Then we have:

THEOREM (2.4). *$N(P, K'')$ is a regular neighbourhood of P.*

THEOREM (2.5). *Any two regular enlargements in M of the same polyhedron are (piecewise linearly) homeomorphic.*

For the proofs of (2.4), (2.5) see [16, p. 293]. We have altered some of the terms used in [16] so as to emphasize the distinction between a "collapse" and an arbitrary retraction by deformation; also to retain the ordinary meaning of "neighbourhood".

If N is a regular enlargement of P and P collapses into P_0, then N is a regular enlargement of P_0. Therefore it follows from (2.5) that every regular enlargement (in M) of P is homeomorphic to every regular enlargement of P_0. By (2.4) a regular enlargement of a point is an n-element. So we have:

COROLLARY (2.6). *Every regular enlargement of a completely collapsible polyhedron is an n-element.*

We now come to the main lemma.

LEMMA (2.7). *Let M be an n-manifold and let $P \subset \operatorname{int} M$ be an $(m-1)$-dimensional polyhedron $(0 < 2m \leqq n)$ such that the inclusion map $i\colon P \to M$ is homotopic in M to a constant. Then there is an n-element $E \subset \operatorname{int} M$ such that $P \subset \operatorname{int} E$.*

PROOF. Let $C = v * P$ be a cone with P as base ($P \subset C$) and vertex v. Assume that i can be extended to an imbedding $h\colon C \to \operatorname{int} M$ and let K be a triangulation of the pair (M, hC). Obviously C, hC are completely collapsible and (2.7), with $E = N(hC, K'')$, follows from (2.4), (2.6). We proceed to prove the existence of h.

Since $i \cong$ const. it can be extended to a (piecewise linear) map $f\colon C \to \operatorname{int} M$. Let P_0 be a triangulation of P and let C_0 be the triangulation of C which consist of the simplexes $v * \sigma$ and their faces, for every simplex $\sigma \in P_0$. We describe f as *normal* if, and only if, it is an imbedding in case $2m < n$ and satisfies the following condition if $2m = n$. If $fx = fy$, where $x, y \in C$, $x \neq y$, then each of x, y is interior to an m-simplex of C_0 and $f^{-1}fx$ contains no point other than x, y. Points such as x, y will be called *singular* (with respect to f).

Assume that f is normal. Then it is an imbedding if $2m < n$. So we assume that $2m = n$. Let $x, y \in C$ be such that $x \neq y$, $fx = fy$ and let

$x \in \text{int}\,(\sigma^m)$, where σ^m is an m-simplex of C_0. If $m = 1$ and σ^m contains more than one singular point let x be the one nearest to $P \cap \sigma^m$. Let $\sigma^{m-1} = P \cap \sigma^m$ and let A be a 1-element in σ^m which joins x to a point $z \in \text{int}\,(\sigma^{m-1})$ and is such that int A does not meet $\dot{\sigma}^m$ or the set of singular points. We shall show how the singular points x, y can be eliminated in the way indicated by the diagram.

Let C_1, M_1 be triangulations of C, M with respect to which f is simplicial, C_1 being a triangulation of the pair (C, A) and a subdivision of C_0. Let C_1'', M_1'' be the second barycentric subdivisions of C_1, M_1 and let P_1'' be the subcomplex of C_1'' which subdivides P_0. Clearly x, y, z are vertices of C_1 and f is simplicial with respect to C_1'', M_1''. Let

$$E_0^m = N(A, C_1'') , \qquad E_1^m = N(y, C_1'') , \qquad E^n = N(fA, M_1'') .$$

The maps $f \mid A$, $f \mid E_i^m$ are imbeddings and it follows from (2.6) that E_i^m, hence also fE_i^m, and E^n are elements. Moreover $E_0^m \cap P_1''$ is the $(m-1)$-element $N(z, P_1'')$. Let $E^{m-1} = \dot{E}_0^m - \text{int}\, N(z, P_1'')$. Then E^{m-1} is an $(m-1)$-element, $fE_i^m \subset E^n$ and

$$\dot{E}^n \cap fE_0^m = fE^{m-1} , \qquad \dot{E}^n \cap fE_1^m = f\dot{E}_1^m \subset \dot{E}^n - fE^{m-1} .$$

Let E_1^n be the second barycentric subdivision of the complex E^n and let

$$E_2^n = N(fE_0^m, E_1^n) , \qquad E^{n-1} = E^n \cap \dot{E}_2^n = N(fE^{m-1}, \dot{E}_1^n) .$$

Then E_2^n, E^{n-1} are regular neighbourhoods in E^n, $\dot{E}^n$ of fE_0^m, fE^{m-1}. Therefore they are elements. Let

$$E_3^n = E^n - \text{int}\, E_2^n - \text{int}\, E^{n-1} \subset E^n - fE_0^m .$$

Then E_3^n is an n-element, by (2.2), and $f\dot{E}_1^m \subset \dot{E}_3^n$. Therefore $f \mid \dot{E}_1^m$ can be extended to an imbedding g: $E_1^m \to E_3^n$. Define f_1: $C \to \text{int}\, M$ by $f_1 = f$ in $C - E_1^m$, $f_1 = g$ in E_1^m. The points x, y are non-singular with respect to f_1 and no new singular points have been introduced. The other singular points, if any, can be eliminated in the same way.

It remains to prove that i has a normal extension. Let $f\colon C \to \operatorname{int} M$ be any (piecewise linear) extension of i. Let $\{E_j\}$ be a finite set of n-elements in M whose interiors cover $\operatorname{int} M$ and let $\varphi_j\colon E_j \to \Delta$ be a homeomorphism on an n-simplex Δ. Let T be a subdivision of C_0 such that, for each simplex $\sigma \in T$ and some $j(\sigma)$,

$$f \operatorname{St}(\sigma, T) \subset \operatorname{int} E_{j(\sigma)} \tag{2.8}$$

(f need not be simplicial with respect to T). Let W be the union of P, T^{k-1} and, possibly, some k-simplexes of T not in P ($0 \leqq k \leqq m$, $T^{-1} = \varnothing$). Assume that $f \mid W$ is normal in the above sense, with W, regarded as a subcomplex of T, playing the part of C_0.

Let τ be a k-simplex of T not in W and define

$$\psi\colon f^{-1}E_j \to \Delta \qquad (j = j(\tau))$$

by $\psi(x) = \varphi_j f(x)$. Since E_j is a polyhedron and f, φ_j are piecewise linear it follows that $f^{-1}E_j$ is a polyhedron and that ψ is piecewise linear. Therefore there is a triangulation K_j of $f^{-1}E_j$, such that ψ is barycentric in each simplex of K_j. Moreover we may assume that K_j is a subcomplex of a subdivision of T. Then it contains a subcomplex which subdivides $\operatorname{St}(\tau, T)$ and every simplex of K_j is contained in a simplex of T. We also assume that K_j has at least one vertex in $\operatorname{int}(\tau)$. Let $b_1, \cdots, b_q$ be the vertices of K_j, ordered so that $b_1, \cdots, b_p$ are the ones in $\operatorname{int}(\tau)$. Let $c_1, \cdots, c_p$ $\operatorname{int}\Delta$ be points which are in general position with respect to each other and to $\psi b_{p+1}, \cdots, \psi b_q$. Let $\psi_1\colon f^{-1}E_j \to \Delta$ be the map, barycentric in each simplex of K_j, which is defined by $\psi_1 b_i = c_i$ or ψb_i according as $i \leqq p$ or $i > p$. If $k = m$ let $c_1, \cdots, c_p$ be such that no m-simplex of $\psi_1 K_j$ with one or more of $c_1, \cdots, c_p$ among its vertices contains a point $\psi_1\sigma_1 \cap \psi_1\sigma_2$, where σ_1 and σ_2 are disjoint closed m-simplexes of $K_j \cap (W \cup \tau)$. Define $f_1\colon C \to \operatorname{int} M$ by

$$\begin{aligned} f_1 x &= \varphi_j^{-1}\psi_1 x && \text{if } x \in f^{-1}E_j \\ &= fx && \text{if } x \in C - \operatorname{int} \operatorname{St}(\tau, T)\,. \end{aligned}$$

The map $f_1 \mid W \cup \tau$ is normal and we take c_i so near to ψb_i ($i = 1, \cdots, p$) that f_1 satisfies (2.8). Then it follows inductively that i has a normal extension, $C \to \operatorname{int} M$, and the proof is complete.

LEMMA (2.9). *Let M, P be as in* (2.7) *and let M be connected and bounded. Then M can be imbedded in $M - P$.*

PROOF. Let E be as in (2.7) and let A be a 1-element which joins a point $x \in \dot{M}$ to a point in $\dot{E}$ and does not meet $\dot{M} \cup E$ anywhere else. Let K be a triangulation of the pair $(M, A \cup E)$ and let $E_0 = N(A \cup E, K'')$. Clear-

ly $A \cup E$ is completely collapsible. Therefore E_0 is an n-element and $\dot{M} \cap E_0$ is the $(n-1)$-element $N(x, K'')$. By (2.2), M is homeomorphic to $M - \text{int } E_0 - \text{int } N(x, K'')$, which is in $M - P$.

ST. JOHN'S COLLEGE, CAMBRIDGE AND PHYSICS DEPT., SYRACUSE UNIVERSITY
MAGDALEN COLLEGE, OXFORD
GONVILLE AND CAIUS COLLEGE, CAMBRIDGE

REFERENCES

1. J. W. ALEXANDER, *The combinatorial theory of complexes*, Ann. of Math., 31 (1930), 292-320.
2. A. L. BLAKERS and W. S. MASSEY, *The homotopy groups of a triad* (II), Ann. of Math., 55 (1952), 192-201.
3. SAMUEL EILENBERG and NORMAN E. STEENROD, Foundations of Algebraic Topology, Princeton, 1952.
4. A. I. FLORES, *Über die Existenz n-dimensionaler Komplexe, die nicht in den R_{2n} einbettbar sind*, Ergebn. Math. Kolloqu., 5 (1933), 17-24.
5. R. H. FOX and J. W. MILNOR, *Singularities of 2-spheres in 4-space and equivalence of knots*, Bull. Amer. Math. Soc., 63 (1957), 406.
6. M. W. HIRSCH, *Immersions of manifolds*, Trans. Amer. Math. Soc., 93 (1959), 242-276.
7. ———, *On imbedding differentiable manifolds in euclidean space*, Ann. of Math., 73 (1961), 566-571.
8. I. M. JAMES, *Some embeddings of projective spaces*, Proc. Camb. Phil. Soc., 55(1959), 294-298.
9. E. R. VAN KAMPEN, *Komplexe in euklidischen Räumen*, Abh. Math. Sem., Univ. Hamburg, 9 (1932), 72-78; 152-153.
10. J. W. MILNOR, *On manifolds homeomorphic to the 7-sphere*, Ann. of Math., 64 (1956), 399-405.
11. M. H. A. NEWMAN, *On the foundations of combinatory analysis situs*, Proc. Kon. Akad. v. Wetensch. Amsterdam, 29 (1926), 627-641.
12. H. SEIFERT and W. THRELFALL, Lehrbuch der Topologie, Leipzig, 1934.
13. ARNOLD SHAPIRO, *Obstructions to the imbedding of a complex in a euclidean space*: I, Ann. of Math., 66 (1957), 256-269.
14. RENÉ THOM, *Espaces fibrés en sphères et carrés de Steenrod*, Ann. Sci. Ecole. Norm. Sup., 69 (1952), 109-182.
15. J. H. C. WHITEHEAD, *On subdivisions of complexes*, Proc. Camb. Phil. Soc., 31 (1935), 69-75.
16. ———, *Simplicial spaces, nuclei and m-groups*, Proc. London Math. Soc., 45 (1939), 243-327.
17. ———, *On C^1-complexes*, Ann. of Math., 41 (1940), 809-824.
18. ———, *The immersion of an open 3-manifold in euclidean 3-space*, to appear in Proc. London Math. Soc..
19. HASSLER WHITNEY, *The self-intersections of a smooth n-manifold in 2n-space*, Ann. of Math., 45 (1944), 220-246.
20. WEN-TSÜN WU, *On the realization of complexes in euclidean spaces*, I, II, Acta Math. Sinica, 5 (1955), 505-552; 7 (1957), 79-101.
21. ——— and GEORGES REEB, Sur les espaces fibrés et les variétés feuilletées, Paris, 1952.
22. E. C. ZEEMAN, *Unknotting spheres*, Ann. of Math., 72 (1960), 350-361.

14
GENERAL RELATIVITY IN SPINOR FORM

1962 *Les theories relativistes de la gravitation* (Royaumont, 1959) Eds. A. Lichnerowicz and M. A. Tonnelat. pp. 429–432. *Éditions du Centre National de la Recherche Scientifique*, Paris.

I was enabled to give this talk at the *International Conference on General Relativity and Gravitation (GR2)*, at Royaumont, France, in June 1958, through the generosity of Dennis Sciama, who gave up half of his own allotted time as an Invited Speaker in order that I could have the opportunity to present my work at this meeting.

GENERAL RELATIVITY IN SPINOR FORM

by Dr R. PENROSE

St John's College, Cambridge

RESUME

Le calcul spinoriel démontre d'une manière frappante plusieurs analogies entre le champ- électromagnétique et le champ du tenseur de Riemann de l'espace-temps vide. Le tenseur de Riemann est représenté par un spineur complètement symétrique à quatre indices, Ψ_{ABCD}. Le produit de ce spineur avec son conjugué complexe donne le tenseur de Robinson-Bel, dont les propriétés ressortent directement de cette formulation. Ψ_{ABCD} est le produit symétrisé de quatre spineurs à un indice. Ainsi il définit quatre vecteurs isotropes. Les différentes manières de coïncidence de ces vecteurs donnent la classification naturelle du tenseur Riemannien. Les dérivées symétrisées de Ψ_{ABCD} sont algébriquement indépendantes et donnent une base algébrique pour toutes les dérivées du tenseur Riemannien.

I propose to show that if a spinor calculus is used in general relativity, instead of the usual tensor calculus, many of the important expressions of the theory take on a surprisingly simple form.

We know from Dirac's theory of the electron that spinors have a fundamental importance to physics in the small, perhaps more so than vectors. I hope to indicate that this may also betrue for the large scale phenomena of gravitation.

Any tensor expression can be translated into a spinor form, each four-dimensional tensor index being replaced by a pair of two-dimensional spinor indices, one undotted and one dotted. This is done by means of a quantity $\sigma_\mu^{A\dot{B}}$ defined at each point of space-time satisfying

$$g^{\mu\nu}\,\sigma_\mu^{A\dot{B}}\,\sigma_\nu^{C\dot{D}} = \epsilon^{AC}\,\epsilon^{\dot{B}\dot{D}},$$

where $\sigma_0^{A\dot{B}}, \sigma_1^{A\dot{B}}, \sigma_2^{A\dot{B}}, \sigma_3^{A\dot{B}}$ are 2×2 hermitian matrices. The spinors ϵ^{AC}, $\epsilon^{\dot{B}\dot{D}}$ (and ϵ_{AC}, $\epsilon_{\dot{B}\dot{D}}$) are the two-dimensional alternating expressions with components 0,1,—1,0 and they determine the skew « metric » of the two spaces (« undotted » and « dotted »). They are used for raising and lowering spinor indices, e.g.

$$\xi^A = \epsilon^{AB}\,\xi_B, \qquad \xi_B = \xi^A\,\epsilon_{AB}\,.$$

The correspondence between tensors and spinors is given by

$$X_\lambda^{\mu\nu} \leftrightarrow X_{A\dot{B}}^{C\dot{D}E\dot{F}} = \sigma_{A\dot{B}}^\lambda\, X_\lambda^{\mu\nu}\,\sigma_\mu^{C\dot{D}}\,\sigma_\nu^{E\dot{F}}\,.$$

When the complex conjugate of a spinor is formed, undotted indices become dotted and dotted indices become undotted. Thus

$$\overline{X}^{\mu\nu}_{\lambda} \leftrightarrow \overline{X}^{\dot{C}\dot{D}\dot{E}\dot{F}}_{\dot{A}\dot{B}}$$

so that reality of tensors is expressed as a hermitian property of the corresponding spinors. Using the symbol ∂_μ(or ∂_{AB}) to denote *covariant* derivative we have

$$\partial_\mu g_{\alpha\beta} = 0, \qquad \partial_\mu \sigma_\nu^{A\dot{B}} = 0, \qquad \partial_\mu \varepsilon^{AC} = 0, \qquad \partial_\mu \varepsilon^{\dot{B}\dot{D}} = 0 .$$

The last two of these equation are often not imposed as they preclude the use of phase transformations to generate the electromagnetic field.

Apart from the general correspondence between tensors and spinors given above, there is also a correspondence between the real skew-symmetric tensors $F_{\mu\nu}$ and the symmetric spinors Φ_{AB} :

$$F_{\mu\nu} \leftrightarrow F_{A\dot{B}C\dot{D}} = \frac{1}{2}\{\Phi_{AC}\,\varepsilon_{\dot{B}\dot{D}} + \varepsilon_{AC}\,\overline{\Phi}_{\dot{B}\dot{D}}\}.$$

The first term on the right is the part of $F_{A\dot{B}C\dot{D}}$ which is symmetric in A, C and therefore skew in $\dot{B}\dot{D}$. The second term is the part skew in A, C and therefore symmetric in $\dot{B}\dot{D}$. Any skew pair of spinor indices can be split of as an ε-spinor.

Now let us use a similar procedure for the Riemann tensor, where for simplicity EINSTEIN's free space equation $R^{\rho}_{\nu\rho\sigma} = 0$ will be imposed. In this case it turns on that we obtain

$$R_{\mu\nu\rho\sigma} \leftrightarrow R_{A\dot{E}B\dot{F}C\dot{G}D\dot{H}} = \frac{1}{2}\{\Psi_{ABCD}\,\varepsilon_{\dot{E}\dot{F}}\,\varepsilon_{\dot{G}\dot{H}} + \varepsilon_{AB}\,\varepsilon_{CD}\,\overline{\Psi}_{\dot{E}\dot{G}\dot{F}\dot{H}}\}.$$

where spinor Ψ_{ABCD} is completely symmetric in all four indices. This correspondence between such spinors Ψ_{ABCD} and such tensors $R_{\mu\nu\rho\sigma}$ is one to one. It was first obtained by WITTEN.

The analogies between the MAXWELL and EINSTEIN fields are brought out in a striking way by the spinor formalism. The source-free MAXWELL equations and the Bianchi identities become, respectively,

$$\partial^{A\dot{C}}\,\Phi_{AB} = 0 \quad \text{and} \quad \partial^{A\dot{E}}\,\Psi_{ABCD} = 0 .$$

The MAXWELL stress-energy tensor and the ROBINSON-BEL « gravitational density » tensor have the respective spinor forms

$$T_{\mu\nu} \leftrightarrow \frac{1}{2}\Phi_{AB}\,\overline{\Phi}_{\dot{C}\dot{D}}, \qquad \frac{1}{2}T_{\mu\nu\rho\sigma} \leftrightarrow \Psi_{ABCD}\,\overline{\Psi}_{\dot{E}\dot{F}\dot{G}\dot{H}} .$$

The fact that $T_{\mu\nu\rho\sigma}$ is totally symmetric in its tensor indices is immediate from the symmetry of Ψ_{ABCD}. The trace-free nature of $T_{\mu\nu\rho\sigma}$ also follows at once from the symmetry in the individual spinor indices. The fact that $T_{\mu\nu}$ is invariant under duality rotations of the electromagnetic field and that $T_{\mu\nu}$ determines this field up to a duality rotation is obvious in the spinor formalism. A duality rotation of the field tensor $F_{\mu\nu}$ corresponds to a phase transformation $\Phi_{AB} \to e^{i\theta}\,\Phi_{AB}$ of the corresponding spinor. Similarly, the curvature tensor (at a point) also admits

duality rotations corresponding to phase transformation of Ψ_{ABCD}. It is again clear from the spinor formalism that $T_{\mu\nu\rho\sigma}$ determines the curvature tensor up to a duality rotation. It is also not hard to show this way that $T_{\mu\nu\rho\sigma}$ is essentially the only four-index tensor which is quadratic in the curvature tensor and which is invariant under these duality rotations.

Symmetric spinors can be decomposed in the following way :

$$\Phi_{AB} = \eta_{(A}\zeta_{B)}, \qquad \Psi_{ABCD} = \alpha_{(A}\beta_B\gamma_C\delta_{D)} .$$

This is unique apart from scale factors. Now any one index spinor ξ_A determines a null vector $x_\mu \leftrightarrow \xi_A \bar{\xi}_{\dot{B}}$. Thus, the electromagnetic field tensor determines two null directions (as is well known) and the curvature tensor determines four null directions. This last fact leads to a natural characterization of curvature tensors (for empty space) in terms of the various ways in which these null directions can coincide. The diagram

$$\begin{array}{ccccc} & & & & (1111) \\ & & & \swarrow & \\ & & (211) & \rightarrow & (22) \\ & \swarrow & & \swarrow & \\ (31) & \rightarrow & (4) & \rightarrow & (—) \\ \text{Petrov type :} & \text{III} & & \text{II} & \text{I} \end{array}$$

shows how the various cases of coincidences can arise from one another as special cases. The symbol (—) denotes the case when the curvature tensor vanishes. All the symbols in any one column correspond to the same Petrov type and those in any one row correspond of the same conditions imposed on the invariants of the curvature tensor. (This classification is, I believe, similar to one obtained by Géhéniau based on some work of Ruse.)

A formula of interest is

$$\Box \Psi_{ABCD} = 3\, \Psi_{(AB}{}^{EF}\Psi_{CD)\,EF} .$$

The non-linear term on the right acts as a kind of « source » term for the Ψ-field. If, instead of empty space we have an electromagnetic field present, the Bianchi identities become

$$\partial^{A}{}_{\dot{E}}\, \Psi_{ABCD} = \frac{k}{2}\, \bar{\Phi}_{\dot{E}\dot{F}}\, \partial_B{}^{\dot{F}}\, \Phi_{CD}$$

where Ψ_{ABCD} is defined from the Weyl tensor. Thus, the Φ-field here acts as a kind of source for the Ψ-field in the first order equation.

Another place in which the spinor formalism leads to a simplification is in picking out a set of algebraically independent expressions from which the curvature tensor and all its derivatives may be obtained algebraically. Such a set is given by

$$\Psi_{AB\cdots G}{}^{\dot{P}\ldots\dot{R}} = \partial_{(E}{}^{(\dot{P}} \cdots \partial_G^{\dot{R})}\, \Psi_{ABCD)} .$$

Thus, for an analytic manifold, these $\Psi_{\ldots\ldots}{}^{\cdots}$ can be specified arbi-

trarily at a point and the curvature at any other point can be calculated from them. As a simple example,

$$\Psi_{AB\cdots G}{}^{\dot{P}\ldots\dot{R}} = a_r \, \Pi_A \, \Pi_B \ldots \Pi_G \, \Pi^{\dot{P}} \ldots \Pi^{\dot{R}}$$

gives plane waves.

DISCUSSION

Intervention du professeur J. Wheeler

Vous avez montré combien de simplifications résultaient de l'emploi du formalisme spinoriel dans l'écriture séparée des équations d'Einstein et de celles de Maxwell. Que se passe-t-il lorsque ces deux groupes d'équations du deuxième ordre sont réunis (Rainich) de manière à donner une théorie du champ purement géométrique et « déjà unifié » (1). Avez-vous essayé d'expliciter les équations du 4e ordre de cette théorie en cette notation spinorielle que vous avez magnifiquement décrite ?

Réponse du docteur Penrose

Je pense que L. Witten a effectué quelques recherches dans cette voie.

(1) C. W. Misner et J. A. Wheeler, *Ann. Phys.*, **2**, 525 (1957).

Intervention du professeur Belinfante

Comment représentez-vous les transformations *générales* des tenseurs par vos transformations spinorielles ? D'habitude on déduit des spineurs uniquement des tenseurs soumis simplement aux transformations de Lorentz.

Réponse

Puisque les $\sigma^{\mu}_{A\dot{B}}$ ne sont pas les matrices habituelles de Pauli et que les $g_{A\dot{B}C\dot{D}}$ sont ici fonctions des coordonnées, les relations entre les spineurs et les tenseurs ne sont pas celles habituellement considérées dans une théorie covariante par rapport au groupe de Lorentz.

Réponse de Penrose à quelques questions :

Au Professeur Wheeler, qui s'est enquis de la possibilité de traduire le travail de Rainich en termes de spineurs : Je pense que Witten a fait quelque chose dans cette direction.

Au Professeur Robinson à propos des rotations de dualité : Il est peut être bon de remarquer que, tandis qu'une rotation de dualité conduit à une nouvelle solution des équations dans le cas du champ de Maxwell, il n'en va pas de même avec le champ ψ, à cause des relations non-linéaires auxquelles ce champ satisfait.

15
AN APPROACH TO GRAVITATIONAL RADIATION BY A METHOD OF SPIN COEFFICIENTS

J. Math. Phys. **3** (1962), 566–578 (Errata **4** (1963), 998).
With E. T. Newman

This work initiated my series of collaborations with Ezra (Ted) Newman, and it became a very widely cited paper. Unfortunately a sign error in the equations prevented Ted from being the first to discover that beautiful solution of Einstein's vacuum equations now known to describe a rotating black hole—the *Kerr* solution, found by Roy Kerr in 1963 (Kerr, R.P. (1963) Gravitational field of a spinning mass as an example of algebraically special metrics. *Phys. Rev. Lett.* **11**, 237–8).

An Approach to Gravitational Radiation by a Method of Spin Coefficients*

EZRA NEWMAN
University of Pittsburgh, Pittsburgh, Pennsylvania
AND
ROGER PENROSE†
Syracuse University,‡ Syracuse, New York
(Received September 29, 1961)

A new approach to general relativity by means of a tetrad or spinor formalism is presented. The essential feature of this approach is the consistent use of certain complex linear combinations of Ricci rotation coefficients which give, in effect, the spinor affine connection. It is applied to two problems in radiation theory; a concise proof of a theorem of Goldberg and Sachs and a description of the asymptotic behavior of the Riemann tensor and metric tensor, for outgoing gravitational radiation.

I. INTRODUCTION

IN the study of gravitational radiation, two techniques have recently gained prominence; the tetrad calculus[1–6] and the spinor calculus.[7–9] In the present paper (Secs. II, III,[10] and IV) it is shown how the two techniques can be used to derive a very compact and useful set of equations, which are essentially linear combinations of the equations for the Riemann tensor expressed in terms of either Ricci rotation coefficients or the spinor affine connection. In Sec. V, we give a short proof of a theorem of Goldberg and Sachs[11] to the effect that if in empty space there exists a null geodesic congruence with vanishing shear, then the Riemann tensor of the space must be algebraically specilized (i.e., the Riemann tensor is not Petrov type I nondegenerate).

The last application of our formalism is to the asymptotic behavior of the Riemann tensor and metric tensor in empty space. In Sec. VI a coordinate system and tetrad are built around a hypersurface-orthogonal null-vector field. In Sec. VII, this special coordinate system and tetrad are used to prove essentially the following theorem. If a particular complex (tetrad) component of the Riemann tensor (complex because we are using a complex tetrad system) has an asymptotic behavior $O(r^{-5})$, the other four complex components are, respectively, $O(r^{-4})$, $O(r^{-3})$, $O(r^{-2})$, and $O(r^{-1})$. The last component represents the pure radiation field. Special cases of this theorem have been known for some time.[12,13] Our theorem is also a slight generalization of a similar result recently obtained by Bondi and Sachs.[6]

II. TETRAD CALCULUS

We deal with a four-dimensional Riemannian space with a signature -2. Into this space a tetrad system of vectors l_μ, m_μ, $\bar{m}_\mu$, n_μ is introduced, l_μ and n_μ being real null vectors and m_μ with its complex conjugate $\bar{m}_\mu$ being complex null vectors. The vector m_μ can be defined from a pair of real, orthogonal unit space-like vectors a_μ and b_μ by $m_\mu = (1/\sqrt{2})(a_\mu - ib_\mu)$. The orthogonality properties of the vectors are

$$\begin{aligned} l_\mu l^\mu = m_\mu m^\mu = \bar{m}_\mu \bar{m}^\mu = n_\mu n^\mu = 0, \\ l_\mu n^\mu = -m_\mu \bar{m}^\mu = 1, \\ l_\mu m^\mu = l_\mu \bar{m}^\mu = n_\mu m^\mu = n_\mu \bar{m}^\mu = 0. \end{aligned} \tag{2.1}$$

It is of great convenience to introduce the tetrad notation[14]

$$z_{\mathbf{m}\mu} = (l_\mu, n_\mu, m_\mu, \bar{m}_\mu), \qquad \mathbf{m} = 1, 2, 3, 4.$$

* This work was done while both authors were at Syracuse University. It was supported by the Aeronautical Research Laboratory.
† Visiting NATO Fellow.
‡ Present address: Mathematics Department, King's College, University of London, London, England.
[1] A. Z. Petrov, Sci. Not. Kazan State Univ. **114**, 55 (1954).
[2] F. A. E. Pirani, Acta Phys. Polon. **15**, 389 (1956); Phys. Rev. **105**, 1089 (1957); Bull. acad. polon. sci. **5**, 143 (1957a).
[3] E. Newman, J. Math. Phys. **2**, 324 (1961).
[4] J. Goldberg and R. Kerr, J. Math. Phys. **2**, 327 (1961).
[5] R. Sachs, Infeld Volume and preprints.
[6] H. Bondi and R. Sachs (private communication).
[7] L. Witten, Phys. Rev. **113**, 357 (1959).
[8] R. Penrose, Ann. Phys. (New York) **10**, 171 (1960).
[9] J. Ehlers, Hamburg Lectures.
[10] It is possible, if one has no familiarity with spinors to omit Sec. III, with but a small loss of continuity.
[11] J. Goldberg and R. Sachs (to be published).
[12] R. Sachs (to be published).
[13] I. Robinson and A. Trautman, Phys. Rev. Letters **4**, 431 (1960).
[14] Greek indices (values 1, 2, 3, 4) are tensor indices, bold face **a**, **b** ··· (values 1, 2, 3, 4) are tetrad indices, capital latin A, B ··· (values 0, 1) are spinor indices and small lightface latin *a*, *b* ··· (values 0, 1) are spinor "dyad" indices.

The tetrad indices can be raised or lowered by the flat-space metric $\eta_{\mathbf{mn}}$

$$\eta_{\mathbf{mn}} = \eta^{\mathbf{mn}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix}. \tag{2.2}$$

The following two relations are easily seen to be true;

$$g_{\mu\nu} = z_{\mathbf{m}\mu} z_{\mathbf{n}\nu} \eta^{\mathbf{mn}}$$

$$= l_\mu n_\nu + n_\mu l_\nu - m_\mu \bar{m}_\nu - \bar{m}_\mu m_\nu \tag{2.3a}$$

$$\eta_{\mathbf{mn}} = z_{\mathbf{m}\mu} z_{\mathbf{n}\nu} g^{\mu\nu}. \tag{2.3b}$$

Complex Ricci rotation coefficients are defined by

$$\gamma_{\mathbf{m}}{}^{\mathbf{np}} = z_{\mathbf{m}\mu;\nu} z^{\mathbf{n}\mu} z^{\mathbf{p}\nu}, \tag{2.4}$$

with the symmetry

$$\gamma^{\mathbf{mnp}} = -\gamma^{\mathbf{nmp}}.$$

The tetrad components of the Riemann tensor defined by

$$R_{\mathbf{mnpq}} = R_{\alpha\beta\gamma\delta} z^{\alpha}_{\mathbf{m}} z^{\beta}_{\mathbf{n}} z^{\gamma}_{\mathbf{p}} z^{\delta}_{\mathbf{q}} \tag{2.5}$$

can be expressed in terms of the rotation coefficients by[15]

$$R^{\mathbf{mnpq}} = \gamma^{\mathbf{mnp};\mathbf{q}} - \gamma^{\mathbf{mnq};\mathbf{p}} + \gamma_{\mathbf{l}}{}^{\mathbf{mq}} \gamma^{\mathbf{lnp}}$$

$$- \gamma_{\mathbf{l}}{}^{\mathbf{mp}} \gamma^{\mathbf{lnq}} + \gamma^{\mathbf{mnl}} (\gamma_{\mathbf{l}}{}^{\mathbf{pq}} - \gamma_{\mathbf{l}}{}^{\mathbf{qp}}), \tag{2.6}$$

with

$$\gamma^{\mathbf{mnp};\mathbf{q}} \equiv \gamma^{\mathbf{mnp}}_{;\mu} z^{\mathbf{q}\mu}.$$

This can easily be derived from the Ricci identity

$$z_{\mathbf{m}\mu;[\alpha\beta]} = \tfrac{1}{2} z_{\mathbf{m}\nu} R^{\nu}_{\mu\alpha\beta} \tag{2.7}$$

by repeated application of (2.4).

The relationship between the Riemann tensor, Weyl tensor, and Ricci tensor goes over in tetrad form unchanged[15]

$$R_{\mathbf{mnpq}} = C_{\mathbf{mnpq}} - \tfrac{1}{2}(\eta_{\mathbf{mp}} R_{\mathbf{nq}} - \eta_{\mathbf{mq}} R_{\mathbf{np}} + \eta_{\mathbf{nq}} R_{\mathbf{mp}}$$

$$- \eta_{\mathbf{np}} R_{\mathbf{mq}}) - (R/6)(\eta_{\mathbf{mq}} \eta_{\mathbf{np}} - \eta_{\mathbf{mp}} \eta_{\mathbf{nq}}). \tag{2.8}$$

In the tetrad notation the Bianchi identities, $R_{\alpha\beta[\gamma\delta;\mu]} = 0$, take the form

$$R_{\mathbf{mn}[\mathbf{pq};\mathbf{r}]} = \gamma_{\mathbf{m}}{}^{\mathbf{l}}{}_{[\mathbf{r}} R_{\mathbf{pq}]\mathbf{ln}}$$

$$- \gamma_{\mathbf{n}}{}^{\mathbf{l}}{}_{[\mathbf{r}} R_{\mathbf{pq}]\mathbf{lm}} + 2 R_{\mathbf{mnl}[\mathbf{p}} \gamma_{\mathbf{r}}{}^{\mathbf{l}}{}_{\mathbf{q}]}. \tag{2.9}$$

Though they appear to be considerably more complicated, it will be seen in Sec. IV that with a new notation they take a simple and useful form in empty space.

In Eq. (2.6) we used the intrinsic derivative defined by $\varphi^{;\mathbf{m}} = \varphi_{;\mu} z^{\mathbf{m}\mu}$. It will be of great value to obtain the commutator of two intrinsic derivatives, $\varphi^{;\mathbf{m};\mathbf{n}} - \varphi^{;\mathbf{n};\mathbf{m}}$. We have

$$\varphi^{;\mathbf{m};\mathbf{n}} = (\varphi_{;\mu} z^{\mathbf{m}\mu})_{;\nu} z^{\mathbf{n}\nu} = \varphi_{;\mu\nu} z^{\mathbf{m}\mu} z^{\mathbf{n}\nu} + \varphi_{;\mu} z^{\mathbf{m}\mu}_{\ ;\nu} z^{\mathbf{n}\nu}. \tag{2.10}$$

By interchanging $\mathbf{m}$ and $\mathbf{n}$ in (2.10) and using $z^{\mathbf{m}\mu}{}_{;\nu} z^{\mathbf{n}\nu} = \gamma^{\mathbf{mpn}} z^{\mu}_{\mathbf{p}}$ obtained from (2.4), we see

$$\varphi^{;\mathbf{m};\mathbf{n}} - \varphi^{;\mathbf{n};\mathbf{m}} = \varphi^{;\mathbf{l}} [\gamma^{\mathbf{m}}{}_{\mathbf{l}}{}^{\mathbf{n}} - \gamma^{\mathbf{n}}{}_{\mathbf{l}}{}^{\mathbf{m}}]. \tag{2.11}$$

In Sec. IV it will be advantageous to dispense with the semicolon notation for intrinsic derivatives and use the following;

$$D\varphi \equiv \varphi_{;\mu} l^{\mu}, \qquad \Delta\varphi \equiv \varphi_{;\mu} n^{\mu}$$

$$\delta\varphi \equiv \varphi_{;\mu} m^{\mu}, \qquad \bar{\delta}\varphi \equiv \varphi_{;\mu} \bar{m}^{\mu}. \tag{2.12}$$

III. TWO-COMPONENT SPINOR CALCULUS

The connection between tensors and spinors[16] is achieved by means of a quantity $\sigma^{\mu}{}_{AB'}$, satisfying

$$g_{\mu\nu} \sigma^{\mu}{}_{AB'} \sigma^{\nu}{}_{CD'} = \epsilon_{AC} \epsilon_{B'D'}. \tag{3.1}$$

For each value of μ, $\sigma_{\mu}{}^{AB'}$ is a (2×2) Hermitian matrix. The ϵ's are Levi-Civita symbols, that is, skew-symmetric expressions with $\epsilon_{01} = \epsilon_{0'1'} = \epsilon^{01} = \epsilon^{0'1'} = 1$, and they are used for lowering or raising spinor indices:

$$\xi^{A} = \epsilon^{AB} \xi_{B}, \qquad \xi_{B} = \xi^{A} \epsilon_{AB},$$

$$\eta^{A'} = \epsilon^{A'B'} \eta_{B'}, \qquad \eta_{B'} = \eta^{A'} \epsilon_{A'B'}. \tag{3.2}$$

(Note the ordering of the indices.) The spinor equivalent of any tensor is a quantity having each tensor index replaced by a pair of spinor indices, one unprimed and one primed[17];

$$X^{\lambda\mu}{}_{\nu} \leftrightarrow X^{AB'CD'}{}_{EF'} = \sigma_{\lambda}{}^{AB'} \sigma_{\mu}{}^{CD'} X^{\lambda\mu}{}_{\nu} \sigma^{\nu}{}_{EF'}.$$

Inversely:

$$X^{\lambda\mu}{}_{\nu} = \sigma^{\lambda}{}_{AB'} \sigma^{\mu}{}_{CD'} X^{AB'CD'}{}_{EF'} \sigma_{\nu}{}^{EF'}.$$

Equation (3.1) tells us that $\epsilon_{AC} \epsilon_{B'D'}$ is the spinor equivalent of $g_{\mu\nu}$.

When taking the complex conjugate of a spinor, unprimed indices become primed, and primed indices become unprimed. For example, the complex conjugate[18] of $X^{AB'CD'}{}_{EF'}$ is $\bar{X}^{A'BC'D}{}_{E'F}$, whence

[15] L. P. Eisenhart, *Riemannian Geometry* (Princeton University Press, Princeton, New Jersey, 1960).

[16] See, for example, W. L. Bade and H. Jehle, Revs. Modern Phys. **25**, 714 (1953).

[17] We use primed rather than dotted indices for typographical reasons.

[18] Many authors omit the bar over the complex conjugate.

the condition for $X^{\lambda\mu}{}_{\nu}$ to be real is the Hermitian property

$$X^{AB'CD'}{}_{EF'} = \bar{X}^{B'AD'C}{}_{F'E}.$$

The covariant derivative $\xi_{A;\mu}$ of the spinor ξ_A is

$$\xi_{A;\mu} = \xi_{A,\mu} - \xi_B \Gamma^B{}_{A\mu}, \tag{3.3}$$

where $\Gamma^B{}_{A\mu}$ is the spinor affine connection. The corresponding quantity $\bar{\Gamma}^{B'}{}_{A'\mu}$ deals with the primed indices. The rules of covariant differentiation for spinor indices are exactly analogous to those for tensor indices. The choice of $\Gamma^B{}_{A\mu}$ is fixed by the requirement that the covariant derivatives of $\sigma^{\mu}{}_{AB'}, \epsilon_{AB}, \epsilon_{A'B'}$ shall all vanish.[16]

Observe that, by (3.1), the four expressions

$$\sigma^{\mu}{}_{00'}, \quad \sigma^{\mu}{}_{11'}, \quad \sigma^{\mu}{}_{01'}, \quad \sigma^{\mu}{}_{10'} \tag{3.4}$$

satisfy the same orthogonality relations (2.1) as the four vectors l^{μ}, n^{μ}, m^{μ}, $\bar{m}^{\mu}$. We would like, therefore, to identify the expressions (3.4) as a convenient tetrad. However, this would not be strictly accurate and is a little misleading. The expressions (3.4) do not really denote vectors as they stand, as is exemplified by the fact that while covariant derivative of $\sigma^{\mu}{}_{AB'}$ is zero, this is not so for l^{μ}, m^{μ}, $\bar{m}^{\mu}$, n^{μ} [see Eq. (2.4)].

To get around this difficulty we introduce two basis spinors o^A, ι^A (a "dyad") normalized thus:

$$o_A \iota^A = \epsilon_{AB} o^A \iota^B = -\iota_A o^A = 1. \tag{3.5}$$

A dyad (in spin space) is analogous to a tetrad in vector space. We may put

$$l^{\mu} = \sigma^{\mu}{}_{AB'} o^A \bar{o}^{B'}, \qquad n^{\mu} = \sigma^{\mu}{}_{AB'} \iota^A \bar{\iota}^{B'},$$
$$m^{\mu} = \sigma^{\mu}{}_{AB'} o^A \bar{\iota}^{B'} \tag{3.6}$$

The covariant derivatives of these expressions will now involve the covariant derivatives of o^A and ι^B.

As with the tetrads, it is convenient to have a generic symbol for both o^A and ι^A. Define $\zeta_a{}^A$, $\bar{\zeta}_a{}^{A'}$ by

$$\zeta_0{}^A = o^A, \qquad \zeta_1{}^A = \iota^A, \qquad \bar{\zeta}_{0'}{}^{A'} = \bar{o}^{A'},$$
$$\bar{\zeta}_{1'}{}^{A'} = \bar{\iota}^{A'}. \tag{3.7}$$

Then, for example, given a spinor $Y_{AB'C}$, we can define its *dyad components*

$$Y_{ab'c} = Y_{AB'C} \zeta_a{}^A \bar{\zeta}_{b'}{}^{B'} \zeta_c{}^C.$$

The lower-case indices behave the same way algebraically as ordinary (capital) spinor indices, but when covariant differentiation is applied, no term involving an affine connection appears for the lower case indices. Thus, the important formal difference between the lower-case and capital indices is simply the difference with respect to covariant differentiation.

Bearing this in mind, it is permissible to choose the components of $\zeta_a{}^A$ to be the Kronecker delta. The dyad components of any spinor will then, in fact, be identical with its spinor components. It is by no means essential to make this specialization but it will be convenient to do so here. The expression

$$\sigma^{\mu}{}_{ab'} = \sigma^{\mu}{}_{AB'} \zeta_a{}^A \bar{\zeta}_{b'}{}^{B'} \tag{3.8}$$

now gives us l^{μ}, n^{μ}, m^{μ}, $\bar{m}^{\mu}$ as ab' take, respectively, the values 00′, 11′, 01′, 10′, by (3.6). With this interpretation, the expressions (3.4) may indeed be though of as giving the required tetrad. However, it is essential to maintain the distinction between the dyad and spinor indices when covariant differentiation is involved.

The components of ζ_{aB} are now the same as those of ϵ_{AB}. Hence, $\zeta_{aA;\mu} = -\zeta_{aC}\Gamma^C{}_{A\mu} = \zeta_a^C \Gamma_{CA\mu} = \Gamma_{aA\mu}$ by (3.3). For the analog of the rotation coefficients[19] (2.4), we therefore have

$$\Gamma_{abcd'} = \zeta_{aA;\mu} \zeta_b{}^A \sigma^{\mu}{}_{cd'} \tag{3.9}$$

with the symmetry

$$\Gamma_{abcd'} = \Gamma_{bacd'}. \tag{3.10}$$

Writing

$$\varphi_{\cdots\cdots;\mu} \sigma^{\mu}{}_{ab'} \equiv \partial_{ab'} \varphi_{\cdots\cdots}, \tag{3.11}$$

the intrinsic derivatives (2.12) become

$$D \equiv \partial_{00'}, \quad \Delta \equiv \partial_{11'}, \quad \delta \equiv \partial_{01'}, \quad \bar{\delta} \equiv \partial_{10'}. \tag{3.12}$$

The commutation relations for these derivatives acting on scalars

$$\{\partial_{ab'} \partial_{cd'} - \partial_{cd'} \partial_{ab'}\}\varphi$$
$$\equiv \{\epsilon^{pq}(\Gamma_{pacd'} \partial_{qb'} - \Gamma_{pcab'} \partial_{qd'})$$
$$+ \epsilon^{p'q'}(\bar{\Gamma}_{p'b'd'c} \partial_{aq'} - \bar{\Gamma}_{p'd'b'a} \partial_{cq'})\}\varphi \tag{3.13}$$

can be obtained by direct calculation from (3.13) using (3.8) and (3.9). By a slight extension of this calculation, when the derivatives act on ζ_{aA} we obtain

$$\partial_{fe'}\Gamma_{acdb'} - \partial_{db'}\Gamma_{acfe'} = \epsilon^{pq}\{\Gamma_{apdb'}\Gamma_{qcfe'}$$
$$+ \Gamma_{acpb'}\Gamma_{qdfe'} - \Gamma_{apfe'}\Gamma_{qcdb'} - \Gamma_{acpe'}\Gamma_{qfdb'}\}$$
$$+ \epsilon^{p'q'}\{\Gamma_{acdr'}\bar{\Gamma}_{q'b'e'f} - \Gamma_{acfr'}\Gamma_{q'e'b'd}\}$$
$$+ \Psi_{acdf}\epsilon_{e'b'} + \Lambda\epsilon_{e'b'}(\epsilon_{cd}\epsilon_{af} + \epsilon_{ab}\epsilon_{ef})$$
$$+ \Phi_{acb'e'}\epsilon_{fd}, \tag{3.14}$$

[19] The quantities (3.9) can be defined directly in terms of derivatives of the $\sigma_{\mu ab'}$, as follows:

$$\Gamma_{abcd'} = \tfrac{1}{2}\epsilon^{p'q'}\{\sigma_{cd'ap'bq'} - \sigma_{cd'bq'ap'} - \sigma_{ap'bq'cd'}\}$$

where

$$\sigma_{ab'cd'ef'} = \sigma^{[\mu}{}_{ab'}\sigma^{\rho]}{}_{cd'}\sigma_{\mu ef'\rho}$$

or

$$\Gamma_{abcd'} = \tfrac{1}{2}\epsilon^{p'q'}\sigma^{\mu}{}_{aq'}\sigma^{\rho}{}_{cd'}\sigma_{\mu bp';\rho}.$$

where the spinors Ψ_{ABCD}, $\Phi_{ABC'D'}$, and Λ correspond, respectively, to the Weyl tensor, trace-free part of the Ricci tensor, and scalar curvature. They have the symmetries:

$$\Psi_{ABCD} = \Psi_{(ABCD)},$$
$$\Phi_{ABC'D'} = \Phi_{(AB)(C'D')} = \bar{\Phi}_{C'D'AB} \tag{3.15}$$

with

$$\Lambda = (1/24)R.$$

The spinor equivalent of the Riemann tensor $R_{\alpha\beta\gamma\delta}$ decomposes as follows[20]:

$$\begin{aligned} -R_{AE'BF'CG'DH'} &= \Psi_{ABCD}\epsilon_{E'F'}\epsilon_{G'H'} \\ &+ \epsilon_{AB}\epsilon_{CD}\bar{\Psi}_{E'F'G'H'} \\ &+ 2\Lambda(\epsilon_{AC}\epsilon_{BD}\epsilon_{E'F'}\epsilon_{G'H'} + \epsilon_{AB}\epsilon_{CD}\epsilon_{E'H'}\epsilon_{F'G'}) \\ &+ \epsilon_{AB}\Phi_{CDE'F'}\epsilon_{G'H'} + \epsilon_{CD}\Phi_{ABG'H'}\epsilon_{E'F'}. \end{aligned} \tag{3.16}$$

The relations[8]

$$\begin{aligned} \partial_{(A}{}^{P'}\partial_{B)P'}\xi_C &= -\Psi_{ABCD}\xi^D + \Lambda\xi_{(A}\epsilon_{B)C}, \\ \partial_{C(P'}\partial^{C}{}_{Q')}\xi_A &= \Phi_{ABP'Q'}\xi^B \end{aligned} \tag{3.17}$$

have been used to obtain (3.14).

The Bianchi identities in spinor form are

$$\begin{aligned} \partial^{D}{}_{G'}\Psi_{ABCD} &= \partial_{(C}{}^{H'}\Phi_{AB)G'H'} \\ \partial^{AG'}\Phi_{ABG'H'} &= -3\,\partial_{BH'}\Lambda \end{aligned} \tag{3.18}$$

from which we obtain

$$\begin{aligned} \partial^{p}{}_{d'}\Psi_{abcp} - \partial_{(c}{}^{t'}\Phi_{ab)d't'} &= \{3\Psi_{pr(ab}\Gamma_{c)}{}^{pr}{}_{d'} \\ &+ \Psi_{abcp}\Gamma^{p}{}_{r}{}^{r}{}_{d'}\} - 2\Gamma^{p}{}_{(ab}{}^{t'}\Phi_{c)pt'd'} \\ &- \{\bar{\Gamma}_{t'd'v'(a}\Phi_{bc)}{}^{t'v'} + \bar{\Gamma}_{t'}{}^{v'}{}_{v'(a}\Phi_{bc)}{}^{t'}{}_{d'}\} \end{aligned} \tag{3.19}$$

and

$$\begin{aligned} &3\,\partial_{ab'}\Lambda + \partial^{pt'}\Phi_{apb't'} \\ &= \epsilon^{v'w'}\{\Phi_{ap}{}^{t'}{}_{w'}\bar{\Gamma}_{b't'v'}{}^{p} + \Phi_{apb'}{}^{t'}\bar{\Gamma}_{t'w'v'}{}^{p}\} \\ &- \{\Phi_{prb'}{}^{t'}\Gamma_{a}{}^{pr}{}_{t'} + \Phi_{apb'}{}^{t'}\Gamma^{p}{}_{r}{}^{r}{}_{t'}\}. \end{aligned} \tag{3.20}$$

IV. THE SPIN COEFFICIENTS

In the present section we will show how the formalisms developed in Secs. II and III can be put into a relatively concise form, despite the fact that all summations will be written out explicitly.

Twelve complex functions (called spin coefficients) are defined in terms of either the rotation coefficients (2.4) or spinor affine connection (3.9).

[20] These definitions of Ψ_{ABCD}, $\Phi_{ABC'D'}$ differ by a factor 2 from those given in reference 8. Also, the Riemann tensor used here is the negative of that used in reference 8.

$$\begin{aligned} \kappa &= \gamma_{131} = l_{\mu;\nu}m^{\mu}l^{\nu}, \quad \pi = -\gamma_{241} = -n_{\mu;\nu}\bar{m}^{\mu}l^{\nu}, \\ \epsilon &= \tfrac{1}{2}(\gamma_{121} - \gamma_{341}) = \tfrac{1}{2}(l_{\mu;\nu}n^{\mu}l^{\nu} - m_{\mu;\nu}\bar{m}^{\mu}l^{\nu}), \\ \rho &= \gamma_{134} = l_{\mu;\nu}m^{\mu}\bar{m}^{\nu}, \quad \lambda = -\gamma_{244} = -n_{\mu;\nu}\bar{m}^{\mu}\bar{m}^{\nu}, \\ \alpha &= \tfrac{1}{2}(\gamma_{124} - \gamma_{344}) = \tfrac{1}{2}(l_{\mu;\nu}n^{\mu}\bar{m}^{\nu} - m_{\mu;\nu}\bar{m}^{\mu}\bar{m}^{\nu}), \\ \sigma &= \gamma_{133} = l_{\mu;\nu}m^{\mu}m^{\nu}, \quad \mu = -\gamma_{243} = -n_{\mu;\nu}\bar{m}^{\mu}m^{\nu}, \\ \beta &= \tfrac{1}{2}(\gamma_{123} - \gamma_{343}) = \tfrac{1}{2}(l_{\mu;\nu}n^{\mu}m^{\nu} - m_{\mu;\nu}\bar{m}^{\mu}m^{\nu}), \\ \nu &= -\gamma_{242} = -n_{\mu;\nu}\bar{m}^{\mu}n^{\nu}, \\ \gamma &= \tfrac{1}{2}(\gamma_{122} - \gamma_{342}) = \tfrac{1}{2}(l_{\mu;\nu}n^{\mu}n^{\nu} - m_{\mu;\nu}\bar{m}^{\mu}n^{\nu}), \\ \tau &= \gamma_{132} = l_{\mu;\nu}m^{\mu}n^{\nu}, \end{aligned} \tag{4.1a}$$

or

$\Gamma_{abcd'} =$

cd' \ ab	00	01 or 10	11
00′	κ	ϵ	π
10′	ρ	α	λ
01′	σ	β	μ
11′	τ	γ	ν

. (4.1b)

It is seen that the spin coefficients appear more naturally when dealing with spinors than with tetrad vectors. This fact reappears when Eq. (2.6) is rewritten in terms of these new functions. The equations are rather unattractive until certain linear combinations are taken. These simpler equations are just the ones, (3.14), that arise naturally in the spinor calculus. Equation (3.14) or the appropriate linear combinations of Eq. (2.6) using (2.8), with the notation of (4.1), is

$$\begin{aligned} D\rho - \bar{\delta}\kappa &= (\rho^2 + \sigma\bar{\sigma}) + (\epsilon + \bar{\epsilon})\rho - \bar{\kappa}\tau \\ &- \kappa(3\alpha + \bar{\beta} - \pi) + \Phi_{00} \end{aligned} \tag{4.2a}$$

$$\begin{aligned} D\sigma - \delta\kappa &= (\rho + \bar{\rho})\sigma + (3\epsilon - \bar{\epsilon})\sigma \\ &- (\tau - \bar{\pi} + \bar{\alpha} + 3\beta)\kappa + \Psi_0 \end{aligned} \tag{4.2b}$$

$$\begin{aligned} D\tau - \Delta\kappa &= (\tau + \bar{\pi})\rho + (\bar{\tau} + \pi)\sigma \\ &+ (\epsilon - \bar{\epsilon})\tau - (3\gamma + \bar{\gamma})\kappa + \Psi_1 + \Phi_{01} \end{aligned} \tag{4.2c}$$

$$\begin{aligned} D\alpha - \bar{\delta}\epsilon &= (\rho + \bar{\epsilon} - 2\epsilon)\alpha + \beta\bar{\sigma} - \bar{\beta}\epsilon \\ &- \kappa\lambda - \bar{\kappa}\gamma + (\epsilon + \rho)\pi + \Phi_{10} \end{aligned} \tag{4.2d}$$

$$\begin{aligned} D\beta - \delta\epsilon &= (\alpha + \pi)\sigma + (\bar{\rho} - \bar{\epsilon})\beta \\ &- (\mu + \gamma)\kappa - (\bar{\alpha} - \bar{\pi})\epsilon + \Psi_1 \end{aligned} \tag{4.2e}$$

$$D\gamma - \Delta\epsilon = (\tau + \bar{\pi})\alpha + (\bar{\tau} + \pi)\beta - (\epsilon + \bar{\epsilon})\gamma - (\gamma + \bar{\gamma})\epsilon + \tau\pi - \nu\kappa + \Psi_2 - \Lambda + \Phi_{11} \quad (4.2f)$$

$$D\lambda - \bar{\delta}\pi = (\rho\lambda + \bar{\sigma}\mu) + \pi^2 + (\alpha - \bar{\beta})\pi - \nu\bar{\kappa} - (3\epsilon - \bar{\epsilon})\lambda + \Phi_{20} \quad (4.2g)$$

$$D\mu - \delta\pi = (\bar{\rho}\mu + \sigma\lambda) + \pi\bar{\pi} - (\epsilon + \bar{\epsilon})\mu - \pi(\bar{\alpha} - \beta) - \nu\kappa + \Psi_2 + 2\Lambda \quad (4.2h)$$

$$D\nu - \Delta\pi = (\pi + \bar{\tau})\mu + (\bar{\pi} + \tau)\lambda + (\gamma - \bar{\gamma})\pi - (3\epsilon + \bar{\epsilon})\nu + \Psi_3 + \Phi_{21} \quad (4.2i)$$

$$\Delta\lambda - \bar{\delta}\nu = -(\mu + \bar{\mu})\lambda - (3\gamma - \bar{\gamma})\lambda + (3\alpha + \bar{\beta} + \pi - \bar{\tau})\nu - \Psi_4 \quad (4.2j)$$

$$\delta\rho - \bar{\delta}\sigma = \rho(\bar{\alpha} + \beta) - \sigma(3\alpha - \bar{\beta}) + (\rho - \bar{\rho})\tau + (\mu - \bar{\mu})\kappa - \Psi_1 + \Phi_{01} \quad (4.2k)$$

$$\delta\alpha - \bar{\delta}\beta = (\mu\rho - \lambda\sigma) + \alpha\bar{\alpha} + \beta\bar{\beta} - 2\alpha\beta + \gamma(\rho - \bar{\rho}) + \epsilon(\mu - \bar{\mu}) - \Psi_2 + \Lambda + \Phi_{11} \quad (4.2l)$$

$$\delta\lambda - \bar{\delta}\mu = +(\rho - \bar{\rho})\nu + (\mu - \bar{\mu})\pi + \mu(\alpha + \bar{\beta}) + \lambda(\bar{\alpha} - 3\beta) - \Psi_3 + \Phi_{21} \quad (4.2m)$$

$$\delta\nu - \Delta\mu = (\mu^2 + \lambda\bar{\lambda}) + (\gamma + \bar{\gamma})\mu - \bar{\nu}\pi + [\tau - 3\beta - \bar{\alpha}]\nu + \Phi_{22} \quad (4.2n)$$

$$\delta\gamma - \Delta\beta = (\tau - \bar{\alpha} - \beta)\gamma + \mu\tau - \sigma\nu - \epsilon\bar{\nu} - \beta(\gamma - \bar{\gamma} - \mu) + \alpha\bar{\lambda} + \Phi_{12} \quad (4.2o)$$

$$\delta\tau - \Delta\sigma = (\mu\sigma + \bar{\lambda}\rho) + (\tau + \beta - \bar{\alpha})\tau - (3\gamma - \bar{\gamma})\sigma - \kappa\bar{\nu} + \Phi_{02} \quad (4.2p)$$

$$\Delta\rho - \bar{\delta}\tau = -(\rho\bar{\mu} + \sigma\lambda) + (\bar{\beta} - \alpha - \bar{\tau})\tau + (\gamma + \bar{\gamma})\rho + \nu\kappa - \Psi_2 - 2\Lambda \quad (4.2q)$$

$$\Delta\alpha - \bar{\delta}\gamma = (\rho + \epsilon)\nu - (\tau + \beta)\lambda + (\bar{\gamma} - \bar{\mu})\alpha + (\bar{\beta} - \bar{\tau})\gamma - \Psi_3. \quad (4.2r)$$

The notation for intrinsic derivatives (2.12) has been used. The quantities Ψ_0, Ψ_1, etc., Φ_{00}, etc., and Λ are, respectively, related to components of the Weyl tensor, Ricci tensor, and scalar curvature by the following;

$$\Psi_0 = -C_{1313} = -C_{\alpha\beta\gamma\delta}l^\alpha m^\beta l^\gamma m^\delta = \Psi_{0000} \quad (4.3a)$$

$$\Psi_1 = -C_{1213} = -C_{\alpha\beta\gamma\delta}l^\alpha n^\beta l^\gamma m^\delta = \Psi_{0001}$$

$$\Psi_2 = -\tfrac{1}{2}(C_{1212} + C_{1234}) = -\tfrac{1}{2}C_{\alpha\beta\gamma\delta} \times (l^\alpha n^\beta l^\gamma n^\delta \mp l^\alpha n^\beta m^\gamma \bar{m}^\delta) = \Psi_{0011}$$

$$\Psi_3 = C_{1224} = C_{\alpha\beta\gamma\delta}l^\alpha n^\beta n^\gamma \bar{m}^\delta = \Psi_{0111}$$

$$\Psi_4 = -C_{2424} = -C_{\alpha\beta\gamma\delta}n^\alpha \bar{m}^\beta n^\gamma \bar{m}^\delta = \Psi_{1111}$$

$$\Phi_{00} = -\tfrac{1}{2}R_{11} = \Phi_{000'0'} = \bar{\Phi}_{00},$$

$$\Phi_{11} = -\tfrac{1}{4}(R_{12} + R_{34}) = \Phi_{010'1'} \quad (4.3b)$$

$$\Phi_{01} = -\tfrac{1}{2}R_{13} = \Phi_{000'1'} = \bar{\Phi}_{10},$$

$$\Phi_{12} = -\tfrac{1}{2}R_{23} = \Phi_{011'1'},$$

$$\Phi_{10} = -\tfrac{1}{2}R_{14} = \Phi_{010'0'} = \bar{\Phi}_{01},$$

$$\Phi_{21} = -\tfrac{1}{2}R_{24} = \Phi_{110'1'},$$

$$\Phi_{02} = -\tfrac{1}{2}R_{33} = \Phi_{001'1'} = \bar{\Phi}_{20},$$

$$\Phi_{22} = -\tfrac{1}{2}R_{22} = \Phi_{111'1'},$$

$$\Phi_{20} = -\tfrac{1}{2}R_{44} = \Phi_{110'0'},$$

$$\Lambda = R/24.$$

With the present notation the commutators (2.11) or (3.13) are

$$(\Delta D - D\Delta)\varphi = [(\gamma + \bar{\gamma})D + (\epsilon + \bar{\epsilon})\Delta - (\tau + \bar{\pi})\bar{\delta} - (\bar{\tau} + \pi)\delta]\varphi \quad (4.4)$$

$$(\delta D - D\delta)\varphi = [(\bar{\alpha} + \beta - \bar{\pi})D + \kappa\Delta - \sigma\bar{\delta} - (\bar{\rho} + \epsilon - \bar{\epsilon})\delta]\varphi$$

$$(\delta\Delta - \Delta\delta)\varphi = [-\bar{\nu}D + (\tau - \bar{\alpha} - \beta)\Delta + \bar{\lambda}\bar{\delta} + (\mu - \gamma + \bar{\gamma})\delta]\varphi$$

$$(\bar{\delta}\delta - \delta\bar{\delta})\varphi = [(\bar{\mu} - \mu)D + (\bar{\rho} - \rho)\Delta - (\bar{\alpha} - \beta)\bar{\delta} - (\bar{\beta} - \alpha)\delta]\varphi.$$

The Bianchi identities (2.9) or (3.18) when written out in general, are very long and unwieldy. However in empty-space, $R_{\alpha\beta} = \Phi_{ABC'D'} = 0$, they do have the simple form[21]

$$D\Psi_1 - \bar{\delta}\Psi_0 = -3\kappa\Psi_2 + [2\epsilon + 4\rho]\Psi_1 - [-\pi + 4\alpha]\Psi_0 \quad (4.5)$$

$$D\Psi_2 - \bar{\delta}\Psi_1 = -2\kappa\Psi_3 + 3\rho\Psi_2 - [-2\pi + 2\alpha]\Psi_1 - \lambda\Psi_0$$

$$D\Psi_3 - \bar{\delta}\Psi_2 = -\kappa\Psi_4 - [2\epsilon - 2\rho]\Psi_3 + 3\pi\Psi_2 - 2\lambda\Psi_1$$

$$D\Psi_4 - \bar{\delta}\Psi_3 = -[4\epsilon - \rho]\Psi_4 + [4\pi + 2\alpha]\Psi_3 - 3\lambda\Psi_2$$

$$\Delta\Psi_0 - \delta\Psi_1 = [4\gamma - \mu]\Psi_0 - [4\tau + 2\beta]\Psi_1 + 3\sigma\Psi_2$$

[21] For completeness, though it is never used in this paper, we give in the Appendix the formulas for the Bianchi identities in the presence of a Maxwell field as well as the Maxwell equations using the notation of this section.

$$\Delta\Psi_1 - \delta\Psi_2 = \nu\Psi_0 + [2\gamma - 2\mu]\Psi_1 - 3\tau\Psi_2 + 2\sigma\Psi_3$$

$$\Delta\Psi_2 - \delta\Psi_3 = 2\nu\Psi_1 - 3\mu\Psi_2 + [-2\tau + 2\beta]\Psi_3 + \sigma\Psi_4$$

$$\Delta\Psi_3 - \delta\Psi_4 = 3\nu\Psi_2 - [2\gamma + 4\mu]\Psi_3 + [-\tau + 4\beta]\Psi_4.$$

Before proceeding to applications of the formalism developed here, it is useful to examine briefly the geometric meaning of some of quantities (4.1) and (4.3) which will be used frequently in the remainder of the paper.

The spin coefficient κ is related to the first curvature of the congruence of which l_μ is the tangent vector by the equation

$$l_{\mu;\nu}l^\nu = -\kappa\bar{m}_\mu - \bar{\kappa}m_\mu + (\epsilon + \bar{\epsilon})l_\mu. \tag{4.6}$$

It is easily seen that if $\kappa = 0$, l_μ is tangent to a geodesic. By a change in scale $l_\mu \to \varphi l_\mu$, $\epsilon + \bar{\epsilon}$ can be made zero. In the case of a geodesic with the above choice of scaling for l_μ we have

$$\rho = \tfrac{1}{2}[-l^\mu{}_{;\mu} + i \text{ curl } l_\mu],$$

where

$$\text{curl } l_\mu = (l_{[\mu;\nu]}l^{\mu;\nu})^{\frac{1}{2}} \tag{4.7}$$

and σ is the complex shear of l_μ satisfying[12,22]

$$\sigma\bar{\sigma} = \tfrac{1}{2}[l_{(\mu;\nu)}l^{\mu;\nu} - \tfrac{1}{2}(l^\mu{}_{;\mu})^2].$$

The quantity τ describes how the direction of l_μ changes as we move in the direction n_μ as follows from the equation

$$l_{\mu;\nu}n^\nu = -\tau\bar{m}_\mu - \bar{\tau}m_\mu + (\gamma + \bar{\gamma})l_\mu. \tag{4.8}$$

Again we can make $\gamma + \bar{\gamma}$ zero by the change $l_\mu \to \varphi l_\mu$.

The spin coefficients ν, μ, λ, π are analogous, respectively, to κ, $-\rho$, $-\sigma$, τ, the difference being that the congruence used is given by n_μ instead of l_μ.

If l_μ is taken tangent to a geodesic congruence and we wish to propogate the remainder of the tetrad system parallelly along this congruence, then

$$\kappa = \epsilon = \pi = 0. \tag{4.9}$$

If in addition to being tangent to geodesics, the l_μ are hypersurface orthogonal, that is, proportional to a gradiant field, we have

$$\rho = \bar{\rho}, \tag{4.10}$$

if *equal* to a gradient field

$$\rho = \bar{\rho}, \qquad \tau = \bar{\alpha} + \beta. \tag{4.11}$$

One can understand the meaning of Ψ_0, Ψ_1, Ψ_2, Ψ_3, and Ψ_4 by the following:

Consider the five cases

(a) $\Psi_0 \neq 0$, others zero
(b) $\Psi_1 \neq 0$, others zero
(c) $\Psi_2 \neq 0$, others zero
(d) $\Psi_3 \neq 0$, others zero
(e) $\Psi_4 \neq 0$, others zero.

The Weyl tensor or the tetrad components of the Weyl tensor will have the following algebraic properties in each of the five cases;

(a) Petrov type N (or [4])[8] with propagation vector n_μ,
(b) Petrov type III (or [31]) with propagation vector n_μ,
(c) Petrov type D (or [22]) with propagation vector n_μ and l_μ,
(d) Petrov type III (or [31]) with propagation vector l_μ,
(e) Petrov type N (or [4]) with propagation vector l_μ.

By a propagation vector, we mean a repeated principal null vector.[8]

If in empty space the vector field l_μ satisfies the equation $l_{[\mu}R_{\alpha]\beta\gamma[\delta}l_{\nu]}l^\beta l^\gamma = 0$, then l_μ corresponds to one of the four principal null directions of the Riemann tensor and

$$\Psi_0 = 0. \tag{4.12}$$

If two or more of the principal null directions coincide and are represented by l_μ, they must satisfy $R_{\alpha\beta\gamma[\delta}l_{\mu]}l^\beta l^\gamma = 0$ or

$$\Psi_0 = \Psi_1 = 0. \tag{4.13}$$

(In this case, one refers to the Riemann tensor as being algebraically specialized.)

In the following section it will be shown that in empty space if the l_μ are tangents to a geodetic congruence whose shear σ vanishes, then (4.13) must be satisfied, and conversely.

V. GOLDBERG-SACHS THEOREM

In this section the conciseness attained by the use of spin coefficients will be illustrated by an example. Here and in the remainder of the paper it will be assumed that we are dealing with empty space, i.e.,

$$R_{\alpha\beta} = 0.$$

[22] I. Robinson, J. Math. Phys. **2**, 290 (1961).

First we will prove that if the Riemann tensor is algebraically specialized, having $\Psi_0 = \Psi_1 = 0$, then $\sigma = \kappa = 0$. With these assumptions the pertinent Bianchi identities become

$$3\sigma\Psi_2 = 0,$$

$$-\delta\Psi_2 = -3\tau\Psi_2 + 2\sigma\Psi_3 \tag{5.1a}$$

$$\Delta\Psi_2 - \delta\Psi_3 = -3\mu\Psi_2 + (-2\tau + 2\beta)\Psi_3 + \sigma\Psi_4$$

$$-3\kappa\Psi_2 = 0$$

$$D\Psi_2 = -2\kappa\Psi_3 + 3\rho\Psi_2 \tag{5.1b}$$

$$D\Psi_3 - \bar{\delta}\Psi_2 = -\kappa\Psi_4 - (2\epsilon - 2\rho)\Psi_3 + 3\pi\Psi_2.$$

It is easily seen from this that unless the space is flat, $\sigma = 0$ by (5.1a) and $\kappa = 0$ by (5.1b).

The converse is more difficult to prove. We assume $\sigma = \kappa = 0$ and wish to prove $\Psi_0 = \Psi_1 = 0$. We can, by a transformation of the form $m_\mu \to e^{i\theta}m_\mu$ and by using a suitable scaling of l_μ, set $\epsilon = 0$.

The pertinent Eqs. (4.2) are then

$$D\rho = \rho^2 \tag{4.2a'}$$

$$0 = \Psi_0 \tag{4.2b'}$$

$$D\tau = (\tau + \bar{\pi})\rho + \Psi_1 \tag{4.2c'}$$

$$D\beta = \beta\bar{\rho} + \Psi_1 \tag{4.2e'}$$

$$\delta\rho = \rho(\bar{\alpha} + \beta) + (\rho - \bar{\rho})\tau - \Psi_1. \tag{4.2k'}$$

With the fact that $\Psi_0 = 0$, [Eq. (4.2b')] the needed Bianchi identities and commutator are

$$\delta\Psi_1 = (4\tau + 2\beta)\Psi_1 \tag{5.2}$$

$$D\Psi_1 = 4\rho\Psi_1 \tag{5.3}$$

$$(D\delta - \delta D)\varphi = (\bar{\pi} - \bar{\alpha} - \beta)\, D\varphi + \bar{\rho}\, \delta\varphi. \tag{5.4}$$

There is yet a freedom in the choice of the vector n_μ, the freedom being that of the so-called "null rotations,"

$$\begin{aligned} l_\mu &\to l_\mu \\ m_\mu &\to m_\mu + a l_\mu \\ n_\mu &\to n_\mu + a\bar{m}_\mu + \bar{a}m_\mu + a\bar{a}l_\mu. \end{aligned} \tag{5.5}$$

This rotation of the tetrad does not change l_μ or disturb the relation $\epsilon = 0$. The complex function a can be chosen so that $\tau = 0$. [It is possible to do this only if $\rho \neq 0$. However from (4.2 k') it is easily seen that if $\rho = 0$, then $\Psi_1 = 0$, and our theorem is proved.]

Equations (5.2) and (5.3) are rewritten

$$\begin{aligned} \delta \ln \Psi_1 &= 2\beta \\ D \ln \Psi_1 &= 4\rho. \end{aligned} \tag{5.6}$$

Taking mixed derivatives and subtracting the two expressions, we have

$$(D\delta - \delta D) \ln \Psi_1 = 2\, D\beta - 4\, \delta\rho$$

$$= 2\beta\bar{\rho} - 4\rho(\bar{\alpha} + \beta) + 6\Psi_1 \tag{5.7}$$

after using (4.2 e') and (4.2 b'). The commutator (5.4) with $\varphi = \ln \Psi_1$, and using (5.6) is

$$(D\delta - \delta D) \ln \Psi_1 = 2\beta\bar{\rho} - 4\rho(\bar{\alpha} + \beta) + 4\rho\bar{\pi}. \tag{5.8}$$

Subtracting (5.8) from (5.7) we have

$$\Psi_1 = \tfrac{2}{3}\bar{\pi}\rho.$$

If this is compared with (4.2 c'), $\Psi_1 = -\bar{\pi}\rho$ we have, since ρ is assumed different from zero, $\Psi_1 = \bar{\pi} = 0$. This completes the proof.[23]

VI. SPECIAL COORDINATES

It is always possible, in a hyperbolic Riemannian manifold, to introduce a family of null hypersurfaces $u = \text{const}$, that is,

$$g^{\mu\nu}u_{,\mu}u_{,\nu} = 0. \tag{6.1}$$

The vectors $l^\mu = g^{\mu\nu}u_{,\nu}$ are tangent to the family of null geodesics lying in the hypersurfaces $u = \text{const}$, and satisfy

$$l^\mu{}_{;\nu}l^\nu = 0. \tag{6.2}$$

Robinson and Trautman[13] show that if one chooses as coordinates $u = x^1$ and an affine parameter[24] along the geodesics $r = x^2$, and two coordinates x^3, x^4 that label the geodesics on each surface $u = \text{constant}$, the metric takes the form ($i, j = 3, 4$)

$$g^{\mu\nu} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & g^{22} & g^{23} & g^{24} \\ 0 & g^{23} & & \\ 0 & g^{24} & & g^{ij} \end{bmatrix}. \tag{6.3}$$

(It is not always most convenient to use an affine parameter as x^2. Sachs uses a "luminosity" parameter, $\tilde{r} = 2/l^\mu{}_{;\mu}$ which makes the g^{12} different from unity. However, for our purposes an affine parameter seems simplest.)

With these coordinates the vector l_μ becomes

$$l_\mu = \delta^1_\mu, \qquad l^\mu = \delta^\mu_2. \tag{6.4}$$

---

[23] Though we have not seen all the details of the Goldberg-Sachs proof, we believe our proof to be essentially equivalent, but, due to the conciseness of our notation, much shorter.

[24] An affine parameter is a parameter along the geodesic, such that the equation for the geodesic takes the standard form. See, for example, E. M. Schrödinger, *Expanding Universes* (Cambridge University Press, New York, 1956).

To preserve $l_\mu n^\mu = 1$, and $l_\mu m^\mu = 0$, we have ($i = 3, 4$)

$$\begin{aligned} m^\mu &= \omega\ \delta_2^\mu + \xi^i\ \delta_i^\mu \\ n^\mu &= \delta_1^\mu + U\ \delta_2^\mu + X^i\ \delta_i^\mu. \end{aligned} \tag{6.5}$$

The relation between the tetrad components (6.4) and (6.5) and the metric components (6.3) is

$$\begin{aligned} g^{22} &= 2(U - \omega\bar{\omega}), \\ g^{2i} &= X^i - (\xi^i\bar{\omega} + \bar{\xi}^i\omega), \\ g^{ij} &= -(\xi^i\bar{\xi}^j + \bar{\xi}^i\xi^j), \end{aligned} \tag{6.6}$$

($i, j = 3, 4$). This follows from (2.3a).

There is still complete freedom for the rotation of the tetrad vectors m^μ and n^μ leaving l^μ fixed. This freedom is eliminated by demanding m^μ and n^μ be parallelly propagated along l^μ. This requirement, in addition to the knowledge that l_μ is a gradient field, is stated in terms of the spin coefficients by (see Sec. IV)

$$\kappa = \pi = \epsilon = 0, \quad \rho = \bar{\rho}, \quad \tau = \bar{\alpha} + \beta. \tag{6.7}$$

With these simplifications the commutators (4.4) are

$$\begin{aligned} (\Delta D - D\Delta)\varphi &= (\gamma + \bar{\gamma})\ D\varphi \\ &\quad - \tau\ \bar{\delta}\varphi - \bar{\tau}\ \delta\varphi \\ (\delta D - D\delta)\varphi &= [\bar{\alpha} + \beta]\ D\varphi \\ &\quad - \sigma\ \bar{\delta}\varphi - \rho\ \delta\varphi \\ (\delta\Delta - \Delta\delta)\varphi &= -\bar{\nu}\ D\varphi \\ &\quad + \bar{\lambda}\ \bar{\delta}\varphi + [\mu + \bar{\gamma} - \gamma]\ \delta\varphi \\ (\bar{\delta}\delta - \delta\bar{\delta})\varphi &= (\mu - \bar{\mu})\ D\varphi \\ &\quad - [\bar{\alpha} - \beta]\ \bar{\delta}\varphi - [\bar{\beta} - \alpha]\ \delta\varphi \end{aligned} \tag{6.8}$$

with [using (6.5) and (2.12)]

$$\begin{aligned} D &= \partial/\partial r, \qquad \delta = \omega\ \partial/\partial r + \xi^i\ \partial/\partial x^i \\ \Delta &= U\ \partial/\partial r + \partial/\partial u + X^i\ \partial/\partial x^i. \end{aligned} \tag{6.9}$$

In order to relate the tetrad components (or metric components) and the spin coefficients we replace φ by u, r, and x^i, respectively, in the four commutators. The result of this operation is ($i = 3, 4$)

$$D\xi^i = \rho\xi^i + \sigma\bar{\xi}^i \tag{6.10a}$$

$$D\omega = \rho\omega + \sigma\bar{\omega} - (\bar{\alpha} + \beta) \tag{6.10b}$$

$$DX^i = \tau\bar{\xi}^i + \bar{\tau}\xi^i \tag{6.10c}$$

$$DU = \tau\bar{\omega} + \bar{\tau}\omega - (\gamma + \bar{\gamma}) \tag{6.10d}$$

$$\delta X^i - \Delta\xi^i = (\mu + \bar{\gamma} - \gamma)\xi^i + \bar{\lambda}\bar{\xi}^i \tag{6.10e}$$

$$\delta\bar{\xi}^i - \bar{\delta}\xi^i = (\bar{\beta} - \alpha)\xi^i + (\bar{\alpha} - \beta)\bar{\xi}^i \tag{6.10f}$$

$$\delta\bar{\omega} - \bar{\delta}\omega = (\bar{\beta} - \alpha)\omega + (\bar{\alpha} - \beta)\bar{\omega} + (\mu - \bar{\mu}) \tag{6.10g}$$

$$\delta U - \Delta\omega = (\mu + \bar{\gamma} - \gamma)\omega + \bar{\lambda}\bar{\omega} - \bar{\nu}. \tag{6.10h}$$

We will refer to these as the metric equations.

To conclude this section we will write the Eqs. (4.2) and the Bianchi identities using the conditions (6.7).

$$D\rho = \rho^2 + \sigma\bar{\sigma} \tag{6.11a}$$

$$D\sigma = 2\rho\sigma + \Psi_0 \tag{6.11b}$$

$$D\tau = \tau\rho + \bar{\tau}\sigma + \Psi_1 \tag{6.11c}$$

$$D\alpha = \alpha\rho + \beta\bar{\sigma} \tag{6.11d}$$

$$D\beta = \beta\rho + \alpha\sigma + \Psi_1 \tag{6.11e}$$

$$D\gamma = \tau\alpha + \bar{\tau}\beta + \Psi_2 \tag{6.11f}$$

$$D\lambda = \lambda\rho + \mu\bar{\sigma} \tag{6.11g}$$

$$D\mu = \mu\rho + \lambda\sigma + \Psi_2 \tag{6.11h}$$

$$D\nu = \tau\lambda + \bar{\tau}\mu + \Psi_3 \tag{6.11i}$$

$$\Delta\lambda - \bar{\delta}\nu = 2\alpha\nu + (\bar{\gamma} - 3\gamma - \mu - \bar{\mu})\lambda - \Psi_4 \tag{6.11j}$$

$$\delta\rho - \bar{\delta}\sigma = (\beta + \bar{\alpha})\rho + (\bar{\beta} - 3\alpha)\sigma - \Psi_1 \tag{6.11k}$$

$$\delta\alpha - \bar{\delta}\beta = \mu\rho - \lambda\sigma - 2\alpha\beta + \alpha\bar{\alpha} + \beta\bar{\beta} - \Psi_2 \tag{6.11l}$$

$$\delta\lambda - \bar{\delta}\mu = (\alpha + \bar{\beta})\mu + (\bar{\alpha} - 3\beta)\lambda - \Psi_3 \tag{6.11m}$$

$$\delta\nu - \Delta\mu = \gamma\mu - 2\nu\beta + \bar{\gamma}\mu + \mu^2 + \lambda\bar{\lambda} \tag{6.11n}$$

$$\delta\gamma - \Delta\beta = \tau\mu - \sigma\nu + (\mu - \gamma + \bar{\gamma})\beta + \bar{\lambda}\alpha \tag{6.11o}$$

$$\delta\tau - \Delta\sigma = 2\tau\beta + (\bar{\gamma} + \mu - 3\gamma)\sigma + \bar{\lambda}\rho \tag{6.11p}$$

$$\Delta\rho - \bar{\delta}\tau = (\gamma + \bar{\gamma} - \bar{\mu})\rho - 2\alpha\tau - \lambda\sigma - \Psi_2 \tag{6.11q}$$

$$\begin{aligned} \Delta\alpha - \bar{\delta}\gamma &= \rho\nu - \tau\lambda - \lambda\beta \\ &\quad + (\bar{\gamma} - \gamma - \bar{\mu})\alpha - \Psi_3 \end{aligned} \tag{6.11r}$$

$$D\Psi_1 - \bar{\delta}\Psi_0 = 4\rho\Psi_1 - 4\alpha\Psi_0 \tag{6.12a}$$

$$D\Psi_2 - \bar{\delta}\Psi_1 = 3\rho\Psi_2 - 2\alpha\Psi_1 - \lambda\Psi_0 \tag{6.12b}$$

$$D\Psi_3 - \bar{\delta}\Psi_2 = 2\rho\Psi_3 - 2\lambda\Psi_1 \tag{6.12c}$$

$$D\Psi_4 - \bar{\delta}\Psi_3 = \rho\Psi_4 + 2\alpha\Psi_3 - 3\lambda\Psi_2 \tag{6.12d}$$

$$\begin{aligned} \Delta\Psi_0 - \delta\Psi_1 &= [4\gamma - \mu]\Psi_0 \\ &\quad - [4\tau + 2\beta]\Psi_1 + 3\sigma\Psi_2 \end{aligned} \tag{6.12e}$$

$$\begin{aligned} \Delta\Psi_1 - \delta\Psi_2 &= \nu\Psi_0 + [2\gamma - 2\mu]\Psi_1 \\ &\quad - 3\tau\Psi_2 + 2\sigma\Psi_3 \end{aligned} \tag{6.12f}$$

$$\begin{aligned} \Delta\Psi_2 - \delta\Psi_3 &= 2\nu\Psi_1 - 3\mu\Psi_2 \\ &\quad + [-2\tau + 2\beta]\Psi_3 + \sigma\Psi_4 \end{aligned} \tag{6.12g}$$

$$\Delta\Psi_3 - \delta\Psi_4 = 3\rho\Psi_2 - [2\gamma + 4\mu]\Psi_3 + [-\tau + 4\beta]\Psi_4. \qquad (6.12h)$$

VII. ASYMPTOTIC BEHAVIOR

We shall now investigate the asymptotic behavior of the Riemann tensor, spin coefficients and metric, for a general type of radiative empty space time. In order to do this, it is necessary to impose some condition of approach to flatness at infinity on the space time. This is usually done in terms of the metric tensor, but it is a little more satisfactory to impose restrictions on the Riemann tensor instead, as we shall do here.

The main condition that will be adopted here is[25]

$$\Psi_0 = O(r^{-5}) \qquad (7.1)$$

but a condition

$$D\Psi_0 = O(r^{-6}) \qquad (7.2)$$

on the r derivative of Ψ_0 will also be used. Furthermore, an assumption of "uniform smoothness" will be imposed, that as many as four or three derivatives with respect to x^3, x^4 do not spoil the above dependence:

$$d_i\Psi_0 = O(r^{-5}), \cdots, d_i\, d_j\, d_k\, d_l\Psi_0 = O(r^{-5}) \quad (i, j, k, l = 3, 4)$$
$$d_i\, D\Psi_0 = O(r^{-6}), \cdots, d_i\, d_j\, d_k\, D\Psi_0 = O(r^{-6}) \qquad (7.3)$$

where

$$d_i \equiv \partial/\partial x^i \qquad (i = 3, 4).$$

It will also be assumed that the hypersurfaces u = const are not so chosen that they are "asymptotically cylindrical" or "asymptotically plane." The exact meaning of this condition will be explained later. It means, in effect, merely that certain very special choices of coordinate system are to be ruled out. From these assumptions[26] we shall prove:

$$\Psi_1 = O(r^{-4}), \quad \Psi_2 = O(r^{-3})$$
$$\Psi_3 = O(r^{-2}), \quad \Psi_4 = O(r^{-1}). \qquad (7.4)$$

If, in addition, we were to assume that the Riemann tensor (in tetrad form) could be expanded[27] in negative powers of r, for large r, then (7.4) would tell us that the coefficient of $r^{-n}(n = 1, \cdots, 5)$ has the algebraic form of an empty-space Riemann tensor having the direction l^a as a $(5-n)$-fold principal null direction. Bondi and Sachs[6] have obtained a similar result under somewhat different (and more restrictive) assumptions.

The choice of Ψ_0 as the quantity whose properties are specified in order to characterize the space-time is in accordance with a certain form of characteristic initial value problem. Given a suitable null hypersurface u = const, the function Ψ_0 on this hypersurface constitutes the main part (and sometimes all) of the initial data[28] for gravitation that is required for continuation. This matter will be discussed elsewhere,[29] but it is worth pointing out here that the determination of $\Psi_1, \cdots, \Psi_4$ from Ψ_0 seems to be a natural first step in this continuation problem. Other quantities such as σ or certain metric variables can be used equally as alternative initial data, but the choice of Ψ_0 seems simpler and is apparently the natural analog for gravitation, of certain characteristic initial data that are appropriate for other fields.

The exponent -5 in (7.1) is in agreement with what one would expect from the linear theory of a radiating quadripole. Also, (7.1) holds for a general null hypersurface in Schwarzschild's solution. If the hypersurfaces u = const open out into the future (i.e., they are analogous to the future null cones given by constant advanced time) then the condition (7.1) would be expected to hold for an isolated system in the absence of incoming radiation, as is suggested by the linearized theory. In fact, even incoming radiation of sufficiently curtailed duration would not be expected to affect (7.1), (7.2), or (7.3).

We now proceed to prove (7.4) from our assumptions.[30] The proof of (7.4) for a particular null

---

[25] The meaning of the order symbols used here is that $f(r, u, x^i) = O[g(r)]$ means $|f(r, u, x^i)| < g(r)\, F(u, x^i)$ for some function F independent of r and for all large r, and $f(r, u, x^i) = o[g(r)]$ means

$$\lim_{r\to\infty} \frac{f(r, u, x^i)}{g(r)} = 0 \text{ for each } u, x^i.$$

[26] These assumptions, though stated in terms of a particular coordinate system appear to have a considerable amount of coordinate independence. For example, given a null geodesic with affine parameter r and tangent vector l_μ, if the r parameter of the original coordinate system can be so adjusted that

$$\bar{r} = r + o(r), \quad \bar{l}_\mu = l_\mu + O(r^{-1}),$$

then (7.4) implies that $\bar{\Psi}_0 = O(r^{-5})$ also, where $\bar{\Psi}_0$ is the complex Riemann tensor component associated with $\bar{l}_\mu$. However, additional global assumptions appear to be necessary to ensure that r can always be so chosen.

[27] This may be a fairly strong restriction. It is, of course, stronger than just local analyticity in r since, for example, $r^{-n} \ln r$ cannot be expanded in negative powers of r.

[28] More properly, the quantity $D\Psi_0 - 5\rho\Psi_0$ may be the most significant one to specify on the hypersurface.

[29] R. Penrose (to be published).

[30] The necessity of (7.3) for the deduction of (7.4) and of ruling out the "asymptotically plane" case can be illustrated by considerations of certain plane waves. Plane waves can also be used to show that, for example, a local assumption merely of $\Psi_4 = O(r^{-1})$ or even $R_{\mu\nu\rho\sigma} = O(r^{-1})$ is quite inadequate for obtaining (7.4).

geodesic u, x^2, x^3 = const will depend on (7.1), (7.2), (7.3) only along this geodesic and its neighborhood within the hypersurface u = const. The order of procedure will be to obtain the r dependence of first ρ, σ and then the various x^i derivatives ($i = 3, 4$) of ρ, σ up to third order. Next, α, β, ξ^i, ω, Ψ_1, are similarly obtained, followed by their x^i derivatives. Then λ, μ, Ψ_2, are treated correspondingly, followed by Ψ_3, Ψ_4. The r dependence of the remaining quantities τ, γ, ν, X^i, U and hence g^{ij} may also be obtained if desired.

Writing

$$P = \begin{bmatrix} \rho & \sigma \\ \bar{\sigma} & \rho \end{bmatrix}, \qquad Q = \begin{bmatrix} 0 & \Psi_0 \\ \bar{\Psi}_0 & 0 \end{bmatrix}$$

(6.11a, b) become

$$DP = P^2 + Q. \tag{7.5}$$

This equation may be solved by

$$P = -(DY)Y^{-1}, \tag{7.6}$$

where

$$Y = \begin{bmatrix} y_1 & y_2 \\ \bar{y}_1 & \bar{y}_2 \end{bmatrix} \tag{7.7}$$

is a nonsingular solution (for given P) of

$$DY = -PY \tag{7.8}$$

and so satisfies

$$D^2Y = -QY. \tag{7.9}$$

The asymptotic behavior of the solutions to (7.9) when $\int r\,|\Psi_0|\,dr = O(1)$, is[31]

$$DY = F + o(1) \tag{7.10}$$

$$Y = rF + o(r) \tag{7.11}$$

where F is a constant matrix. We can improve on this here since $Q = O(r^{-5})$. From (7.9) and (7.11) we get

$$D^2Y = -rQF + o(r^{-4}) = O(r^{-4}).$$

Hence, integrating[32] twice and comparing with (7.10), we get

$$DY = F + O(r^{-3}), \tag{7.12}$$

$$Y = rF + E + O(r^{-2}), \tag{7.13}$$

where E is another constant matrix. The solution (7.6) for P can now be used giving

$$P = -r^{-1}I + r^{-2}EF^{-1} + O(r^{-3}) \tag{7.14}$$

provided F is nonsingular. If F is singular, the asymptotic behavior of P is quite different. The case $|F| = 0$, $F \neq 0$, gives the "asymptotically cylindrical" case and P becomes asymptotically proportional to a singular matrix, with $\rho = -\frac{1}{2}r^{-1} + O(r^{-2})$. If $F = 0$, we get the "asymptotically plane" case and $P = O(r^{-3})$. [In each case, E must be such that there are two linearly independent columns among those of E, F. Otherwise it follows that in fact $|Y| = 0$ for all r so we do not get a solution of (7.5) for P.] These two exceptional cases are to be ruled out by assumption. For any given null geodesic u, x^2, x^3 = const the two types of exceptional behavior can always be avoided by an arbitrarily small change in the coordinate system which changes the relevant hypersurface u = const into a slightly different one through this null geodesic. Such exceptional solutions of the Eq. (7.5) do not occur for general choices of initial values for ρ and σ.

From (7.14) we get

$$\rho = -r^{-1} + O(r^{-2}), \qquad \sigma = O(r^{-2}). \tag{7.15}$$

In fact, a lot more can be obtained about the asymptotic behavior of ρ and σ, but (7.15) is all that will be needed here.

In order to proceed further, we shall require the following lemma:

Lemma. Let the complex $(n \times n)$ matrix B and the complex column n vector b be given functions of r where

$$B = O(r^{-2}), \qquad b = O(r^{-2}). \tag{7.16}$$

Let the $(n \times n)$ matrix A be independent of r and have no eigenvalue with a positive real part. Then all of the solution of

$$Dy = (Ar^{-1} + B)y + b \tag{7.17}$$

are bounded as $r \to \infty$, y being a complex column n vector function of r.

Proof. Put $r = e^l$, then (7.17) can be rewritten:

$$\frac{d}{dl}\begin{pmatrix} y \\ 1 \end{pmatrix} = \left\{ \left[\begin{array}{c|c} A & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline 0 \cdots 0 & 0 \end{array}\right] + \left[\begin{array}{c|c} C & c \\ \hline 0 \cdots 0 & 0 \end{array}\right] \right\} \begin{pmatrix} y \\ 1 \end{pmatrix} \tag{7.18}$$

[31] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations* (McGraw-Hill Book Publishers Inc., New York, 1955), p. 103.

[32] It is permissible to integrate order symbols formally but not to differentiate them.

where

$$C(l) = e^{l}B(e^{l}), \qquad c(l) = e^{l}b(e^{l}). \tag{7.19}$$

Now, the solutions ($z = \{\exp Al\}z_0$, $\zeta = \zeta_0$) of

$$\frac{d}{dl}\begin{pmatrix} z \\ \zeta \end{pmatrix} = \left[\begin{array}{ccc|c} & & & 0 \\ & A & & \vdots \\ & & & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}\right]\begin{pmatrix} z \\ \zeta \end{pmatrix}$$

are all bounded. Also, by (7.16) and (7.19) it follows that the integrals with respect to l of the moduli of the elements of C and of c are bounded as $l \to \infty$. Hence by a theorem of N. Levinson,[33] the solutions of (7.18) are also all bounded [with $y = \{\exp Al\}y_0 + O(1)$]. This proves the lemma.[34]

Suppose, now, that B, b, and y are also functions of x^3, x^4. Then, differentiating (7.17) with respect to x^i we get

$$D(d_i y) = (Ar^{-1} + B)(d_i y) + \{(d_i B)y + d_i b\}, \qquad (i = 3, 4)$$

which is again of the form (7.17) provided that

$$d_i B = O(r^{-2}), \qquad d_i b = O(r^{-2}).$$

If this is the case, it follows from the lemma that $d_i y$ must also be bounded. Repeating this, we get the correspondiag results for higher derivatives.

Now consider the x^i derivatives of (6.11a, b) which can be put in the form

$$D\left\{r^2 d_i \begin{pmatrix} \rho \\ \sigma \\ \bar{\tau} \end{pmatrix}\right\} = \begin{pmatrix} 2\rho + 2r^{-1} & \bar{\tau} & \sigma \\ 2\sigma & 2\rho + 2r^{-1} & 0 \\ 2\bar{\tau} & 0 & 2\rho + 2r^{-1} \end{pmatrix} \times r^2 d_i \begin{pmatrix} \rho \\ \sigma \\ \bar{\tau} \end{pmatrix} + r^2 d_i \begin{pmatrix} 0 \\ \Psi_0 \\ \bar{\Psi}_0 \end{pmatrix}. \tag{7.20}$$

The lemma applies with $A = 0$, by (7.15), (7.3), whence

$$d_i\rho = O(r^{-2}), \quad d_i\sigma = O(r^{-2}) \quad (i = 3, 4). \tag{7.21}$$

The lemma applies again to the next two x^i derivatives of (7.20) successively, whence

$$d_i\, d_j\rho,\ d_i\, d_j\sigma,\ d_i\, d_j\, d_k\rho,\ d_i\, d_j\, d_k\sigma = O(r^{-2}). \tag{7.22}$$

Next, using (6.9), we can apply the lemma to (6.12a), (6.11d, e), (6.10a, b), and their complex conjugates, y being the column vector

$$\{r^4\Psi_1, r^4\bar{\Psi}_1, r\alpha, r\bar{\alpha}, r\beta, r\bar{\beta}, r\xi^3, r\bar{\xi}^3, r\xi^4, r\bar{\xi}^4, \omega, \tilde{\omega}\},$$

$b = 0$

and

$$A = \left[\begin{array}{ccc|cc} 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ \hline & -a & & -1 & 0 \\ & & & 0 & -1 \end{array}\right],$$

$$a = \begin{bmatrix} 0&0&0&1&1&0&0&0&0&0 \\ 0&0&1&0&0&1&0&0&0&0 \end{bmatrix},$$

the elements of B being $O(r^{-2})$ expressions in r, ρ, σ, Ψ_0, $D\Psi_0$, and $d_i\Psi_0$. Hence

$$\Psi_1 = O(r^{-4}) \tag{7.23}$$

and

$$\alpha, \beta, \xi^3, \xi^4 = O(r^{-1}), \qquad \omega = 0(1). \tag{7.24}$$

Also, by (6.12a)

$$D\Psi_1 = O(r^{-5}). \tag{7.25}$$

Taking successive x^i derivatives and using (7.21), (7.22), and (7.3), the lemma also gives

$$d_i\Psi_1, \qquad d_i\, d_j\Psi_1, \qquad d_i\, d_j\, d_k\Psi_1 = O(r^{-4}),$$

$$d_i\alpha, \cdots, d_i\xi^4, \qquad d_i\, d_j\alpha, \cdots, d_i\, d_j\, d_k\xi^4 = O(r^{-1}),$$

$$d_i\omega, \cdots, d_i\, d_j\, d_k\omega = O(1)$$

and so by (6.12a)

$$d_i\, D\Psi_1, \qquad d_i\, d_j\, D\Psi_1 = O(r^{-5}).$$

Next consider (6.12b) and (6.11g, h). The lemma applies again with y as the column vector $\{r^3\Psi_2, r\lambda, r\mu\}$, with $A = 0$ and B, b as certain $O(r^{-2})$ expressions in r, ρ, σ, Ψ_0, α, Ψ_1, ω, $D\Psi_1$, ξ^i, $d_i\Psi_1$. Thus

$$\Psi_2 = O(r^{-3}) \tag{7.28}$$

$$\lambda, \mu = O(r^{-1}). \tag{7.29}$$

Hence,

$$D\Psi_2 = O(r^{-4}),$$

[33] N. Levinson, Am. J. Math. **68**, 1 (1946).

[34] It may be seen from the proof to the lemma that conditions (7.16) are in fact, rather stronger than is necessary. They may be weakened to B, $b = O[f(r)]$ where $\int f\,dr = 0(1)$, $f > 0$. This enables condition (7.2) to be weakened to $D\,\Psi_0 = O(r^{-4}f(r))$ and (7.4) can still be obtained. Conditions (7.3) [and even (7.1)] can also be correspondingly weakened.

and continuing with the x^i derivatives we get

$$d_i\Psi_2, \quad d_id_j\Psi_2 = O(r^{-3});$$

$$d_i\lambda, \cdots, d_id_j\mu = O(r^{-1}); d_iD\Psi_2 = O(r^{-4}).$$

Similarly, the lemma applies to (6.12c) with $y = r^2\Psi_3$, then to its x^i derivative and then to (6.12d) with $y = r\Psi_4$ giving

$$\Psi_3 = O(r^{-2}) \tag{7.30}$$

$$d_i\Psi_3 = O(r^{-2}), \qquad D\Psi_3 = O(r^{-3})$$

and then

$$\Psi_4 = O(r^{-1}). \tag{7.31}$$

We may also continue the process and obtain

$$\tau = O(r^{-1}); \quad \gamma, \nu, X^3, X^4 = O(1); \quad U = O(r); \tag{7.32}$$

whence by (6.6)

$$g^{22} = O(r), \qquad g^{2i} = O(1), \qquad g^{ij} = O(r^{-2}),$$
$$(i, j = 3, 4). \tag{7.33}$$

It is possible to obtain a great deal more information about the asymptotic behavior of all these quantities by examining the above procedure a little more closely and then substituting the expressions obtained back into the equations. Also, by specializing the coordinate system further many simplifications can be obtained. (We have, in fact, not even used $\tau = \bar{\alpha} + \beta$ here.) This, together with the integration of the remainder of the Eqs. (6.10) and (6.11), will be discussed elsewhere.

VIII. CONCLUSIONS

In the last section we showed that under certain fairly general assumptions of approach to flatness at infinity that are to be expected in radiative empty spaces, the Riemann tensor exhibits a characteristic asymptotic behavior. This is given by (7.1), (7.23), (7.28), (7.30), and; (7.31); namely

$$\Psi_n = O(r^{-5+n}) \qquad (n = 0, 1, \cdots, 4).$$

We may thus, in general, break the space up into five regions; namely, a near zone, where all terms are important; three transition zones, where Ψ_0, Ψ_1, Ψ_2 become negligible in turn; and finally the radiation zone, where only Ψ_4 remains important and the Riemann tensor is essentially null. The fourth zone is, of course, generally essentially type III and the third is essentially "algebraically special" or usually type II.[5,13] The second zone is essentially a region in which there are "goedesic rays" in the terminology of Sachs[12] and the first zone is of "general" type.[6] Thus, as we move backwards from infinity along a suitable null geodesic the principal null directions "peel off" one by one from the (outgoing) radial direction. This behavior was first observed by Sachs by considering the linear theory.[6] (It must be pointed out, however, that in many particular cases the actual positions of the principal null directions and the Petrov types encountered may not, in fact, agree with the above in detail since some of the Ψ's may be fortuitously small in some regions.)

The analogy between the above and the case of electrodynamics is striking. In the latter case there are three regions to consider, namely, the near zone where r^{-3} terms are important, the transition zone where r^{-2} terms are important and the radiation zone where the field goes essentially as r^{-1} and is null. The two electromagnetic principal null directions exhibit, in the general case, the same characteristic "peeling off" as in the gravitational case.

An interesting further question to consider will be the corresponding "peeling off" theorem for the Einstein-Maxwell theory. The relevant spin coefficient equations are given in an appendix. Another question of importance here is that of the extent to which the assumptions made here are coordinate independent.

ACKNOWLEDGMENTS

We would like to thank the entire relativity group at Syracuse University, in particular E. Schücking and I. Robinson, and also R. K. Sachs King's College, London, for many enlightening discussions.

APPENDIX

In the spin coefficient notation Maxwell's equations take the form

$$\begin{aligned} D\Phi_1 - \bar{\delta}\Phi_0 &= (\pi - 2\alpha)\Phi_0 + 2\rho\Phi_1 - \kappa\Phi_2 \\ D\Phi_2 - \bar{\delta}\Phi_1 &= -\lambda\Phi_0 + 2\pi\Phi_1 + (\rho - 2\epsilon)\Phi_2 \\ \delta\Phi_1 - \Delta\Phi_0 &= (\mu - 2\gamma)\Phi_0 + 2\tau\Phi_1 - \sigma\Phi_2 \\ \delta\Phi_2 - \Delta\Phi_1 &= -\nu\Phi_0 + 2\mu\Phi_1 + (\tau - 2\beta)\Phi_2 \end{aligned} \tag{A1}$$

with

$$\Phi_0 = F_{\mu\nu}l^\mu m^\nu, \qquad \Phi_1 = \tfrac{1}{2}F_{\mu\nu}(l^\mu n^\nu + \bar{m}^\mu m^\nu),$$
$$\Phi_2 = F_{\mu\nu}\bar{m}^\mu n^\nu.$$

If the Ricci tensor is proportional to the Maxwell stress tensor, so that

$$\Phi_{mn} = k\Phi_m\bar{\Phi}_n \qquad (m, n = 0, 1, 2) \tag{A2}$$

then the Bianchi identities become (choosing $k = 1$ for convenience)

$$
\begin{aligned}
&(\delta - \tau + 4\beta)\Psi_4 - (\Delta + 2\gamma + 4\mu)\Psi_3 + 3\nu\Psi_2 \\
&\qquad = \bar{\Phi}_1\,\Delta\Phi_2 - \bar{\Phi}_2\,\bar{\delta}\Phi_2 + 2(\bar{\Phi}_1\Phi_1\nu - \bar{\Phi}_2\Phi_1\lambda - \bar{\Phi}_1\Phi_2\gamma + \bar{\Phi}_2\Phi_2\alpha), \\
&(\delta - 2\tau + 2\beta)\Psi_3 + \sigma\Psi_4 - (\Delta + 3\mu)\Psi_2 + 2\nu\Psi_1 \\
&\qquad = \bar{\Phi}_1\,\delta\Phi_2 - \bar{\Phi}_2\,D\Phi_2 + 2(\bar{\Phi}_1\Phi_1\mu - \bar{\Phi}_2\Phi_1\pi - \bar{\Phi}_1\Phi_2\beta + \bar{\Phi}_2\Phi_2\epsilon), \\
&(\delta - 3\tau)\Psi_2 + 2\sigma\Psi_3 - (\Delta - 2\gamma + 2\mu)\Psi_1 + \nu\Psi_0 \\
&\qquad = \bar{\Phi}_1\,\Delta\Phi_0 - \bar{\Phi}_2\,\delta\Phi_0 + 2(\bar{\Phi}_1\Phi_0\gamma - \bar{\Phi}_2\Phi_0\alpha - \bar{\Phi}_1\Phi_1\tau + \bar{\Phi}_2\Phi_1\rho), \\
&(\delta - 4\tau - 2\beta)\Psi_1 + 3\sigma\Psi_2 - (\Delta - 4\gamma + \mu)\Psi_0 \\
&\qquad = \bar{\Phi}_1\,\delta\Phi_0 - \bar{\Phi}_2\,D\Phi_0 + 2(\bar{\Phi}_1\Phi_0\beta - \bar{\Phi}_2\Phi_0\epsilon - \bar{\Phi}_1\Phi_1\sigma + \bar{\Phi}_2\Phi_1\kappa), \qquad \text{(A3)} \\
&(D + 4\epsilon - \rho)\Psi_4 - (\bar{\delta} + 4\pi + 2\alpha)\Psi_3 + 3\lambda\Psi_2 \\
&\qquad = \bar{\Phi}_0\,\Delta\Phi_2 - \bar{\Phi}_1\,\bar{\delta}\Phi_2 + 2(\bar{\Phi}_0\Phi_1\nu - \bar{\Phi}_1\Phi_1\lambda - \bar{\Phi}_0\Phi_2\gamma + \bar{\Phi}_1\Phi_2\alpha), \\
&(D + 2\epsilon - 2\rho)\Psi_3 + \kappa\Psi_4 - (\bar{\delta} + 3\pi)\Psi_2 + 2\lambda\Psi_1 \\
&\qquad = \bar{\Phi}_0\,\delta\Phi_2 - \bar{\Phi}_1\,D\Phi_2 + 2(\bar{\Phi}_0\Phi_1\mu - \bar{\Phi}_1\Phi_1\pi - \bar{\Phi}_0\Phi_2\beta + \bar{\Phi}_1\Phi_2\epsilon), \\
&(D - 3\rho)\Psi_2 + 2\kappa\Psi_3 - (\bar{\delta} + 2\pi = 2\alpha)\Psi_1 + \lambda\Psi_0 \\
&\qquad = \bar{\Phi}_0\,\Delta\Phi_0 - \bar{\Phi}_1\,\delta\Phi_0 + 2(\bar{\Phi}_0\Phi_0\gamma - \bar{\Phi}_1\Phi_0\alpha - \bar{\Phi}_0\Phi_1\tau + \bar{\Phi}_1\Phi_1\rho), \\
&(D - 2\epsilon - 4\rho)\Psi_1 + 3\kappa\Psi_2 - (\bar{\delta} + \pi - 4\alpha)\Psi_0 \\
&\qquad = \bar{\Phi}_0\,\delta\Phi_0 - \bar{\Phi}_1\,D\Phi_0 + 2(\bar{\Phi}_0\Phi_0\beta - \bar{\Phi}_1\Phi_0\epsilon - \bar{\Phi}_0\Phi_1\sigma + \bar{\Phi}_1\Phi_1\kappa).
\end{aligned}
$$

Errata: An Approach to Gravitational Radiation by a Method of Spin Coefficients

E. NEWMAN
University of Pittsburgh, Pittsburgh, Pennsylvania
AND
R. PENROSE
Kings College, University of London, London, England
[J. Math. Phys. **3**, 566 (1962)]

Page 568: The next-to-the-last term in Eq. (3.14) should read $\Lambda\epsilon_{a'b'}(\epsilon_{cd}\epsilon_{af} + \epsilon_{ad}\epsilon_{cf})$;

Page 569: The last term in the first equation of (3.17) should read $2\Lambda\xi_{(A}\epsilon_{B)C}$;

Page 570: The third term from the end of Eq. (4.2g) should read $-\nu\tilde{\kappa}$;
The third term from the end of Eq. (4.2k) should read $+(\mu - \bar{\mu})\kappa$;
The first term after the equals sign in Eq. (4.2m) should read $+(\rho - \bar{\rho})\nu$;
The third term from the end of Eq. (4.2r) should read $+(\tilde{\gamma} - \bar{\mu})\alpha$;
The third equation of (4.3a) should read $\Psi_2 = -\frac{1}{2}(C_{1212} - C_{1234}) = -\frac{1}{2}C_{\alpha\beta\gamma\delta} \times (l^\alpha n^\beta l^\gamma n^\delta - l^\alpha n^\beta m^\gamma \bar{m}^\delta) = \Psi_{0011}$.

Page 572: The second sentence after Eq. 6.3 should read "Sachs uses a "luminosity" parameter $\tilde{r}$, satisfying $D\tilde{r} = -\tilde{r}\rho$, which $\cdots$," etc.

Page 574: In footnote 26, the second sentence should read, " $\cdots$ affine parameter $\tilde{r}$ and tangent vector $\tilde{l}_\mu$ $\cdots$ ".

Page 575: The lemma should read, "Let the $(n \times n)$ matrix A be independent of r and have no eigenvalue with positive real part. Suppose, also that any eigenvalue of A with vanishing real part is regular (i.e. its multiplicity is equal to the number of linearly independent eigenvectors corresponding to it). Then all the solutions $\cdots$ ".

Page 578: In Eqs. (A3) of the Appendix, the differentiated terms on the right-hand side all appear with the wrong sign. Also the δ's on the right-hand sides of the third, fifth, and seventh equations should be $\bar{\delta}$'s. The equality symbol on the left in the seventh equation should, of course, be a minus sign. As an example this seventh equation should be correctly written

$$(D - 3\rho)\Psi_2 + 2\kappa\Psi_3 - (\bar{\delta} + 2\pi - 2\alpha)\Psi_1 + \lambda\Psi_0 = -\bar{\Phi}_0\Delta\Phi_0 + \bar{\Phi}_1\bar{\delta}\Phi_0 + 2(\bar{\Phi}_0\Phi_0\gamma - \bar{\Phi}_1\Phi_0\alpha - \bar{\Phi}_0\Phi_1\tau + \bar{\Phi}_1\Phi_1\rho).$$

16
ASYMPTOTIC PROPERTIES OF FIELDS AND SPACE-TIMES

Phys. Rev. Letters **10** (1963), 66–68

This rather concentrated letter–article introduced the study of asymptotic properties of a space-time $\mathscr{M}$ from the perspective of its *conformal* (light-cone) structure. A *boundary* $\mathscr{I}$ (referred to as *conformal infinity*) is adjoined smoothly to the conformal manifold $\mathscr{M}$. Later work has indicated that to demand, as here, that a past boundary $\mathscr{I}^-$ and future boundary $\mathscr{I}^+$ can *both* join on to $\mathscr{M}$ with a high degree of smoothness may be too restrictive, if we require full freedom for both incoming and outgoing (gravitational) radiation. However, most calculations are concerned only with the outgoing radiation (gravitational and electromagnetic, say) at $\mathscr{I}^+$, and a lack of significant smoothness at $\mathscr{I}^-$ plays no role. Full freedom in outgoing radiation is compatible with a high degree of differentiability at $\mathscr{I}^+$ if we do not demand the same for $\mathscr{I}^-$. The picture of asymptotic flatness is, in any case, to be taken as only an approximation to reality, the essential requirement being that the assumptions allow full freedom for the required dynamics, including radiation. Then $\mathscr{I}^+$ is sufficient. The letter $\mathscr{I}$ is to be pronounced 'scri' (which stands for 'script I'). Coincidentally, the relativity theorists' pronunciation of 'scri' sounds exactly like a Polish word 'skraj', which means a boundary.

A. Saladin for help with the electrostatic separators.

[1]The fitting program was written by A. G. Wilson of the Rutherford Laboratory, National Institute for Research in Nuclear Science, Harwell, England. We are greatly indebted to Mr. Wilson for his help.

[2]The number rejected for this reason was about 12% of the number finally accepted as three-pion states.

[3]This description in terms of l and L is correct only in the nonrelativistic limit, and so the use of this representation in the present experiment where the ρ meson has a momentum of 780 MeV/c is an approximation.

[4]The lowest $^1P_1$ state leading to ρ^0 production is (0,1), $J=1^+$. It has been discussed in connection with the ω^0 spin analysis [M. L. Stevenson, L. W. Alvarez, B. C. Maglić, and A. H. Rosenfeld, Phys. Rev. 125, 687 (1962)].

[5]It has been suggested [S. Glashow, Phys. Rev. Letters 7, 469 (1961); J. Bernstein and G. Feinberg, Brookhaven National Laboratory report BNL 6122, 1962 (unpublished)] that the rate of decay of ω^0 into $\pi^+ + \pi^-$ by an electromagnetic interaction may be appreciable. The reaction $p+\bar{p} \to \omega^0 + \pi^-$ may proceed from S states only through the $^3S_1$ channel of isotopic spin 1, and so any contribution would be added coherently to the direct $^3S_1$ $\rho^0 + \pi^0$ amplitude. In this case equality of the rates of production of $\rho^0\rho^+\rho^-$ would in general imply some contribution of charged ρ's from, say, the $^1S_0$ state unless the interference in the $^3S_1$ channel was so arranged as to lead to the equality. The angular distribution of the charged ρ decay gives no evidence for charged ρ production from the $^1S_0$ state. Similarly it is not possible to exclude P-state contribution completely since interference between several (L,l) waves may occur in such a way as to allow the effect observed.

[6]C. Bouchiat and G. Flamand, Nuovo Cimento 23, 13 (1962).

[7]The relation used is $\gamma = \Gamma_0(q_\gamma^2 + m_\pi^2)/2q_\gamma^3$. Formula (5a) is a P-wave effective range approximation [B. Lee and M. T. Vaughn, Phys. Rev. Letters 4, 578 (1962)].

[8]D. D. Carmony and R. T. Van de Walle, Phys. Rev. Letters 8, 73 (1962).

[9]The contamination from the pseudo-three-pion events is included in this background

ASYMPTOTIC PROPERTIES OF FIELDS AND SPACE-TIMES*

Roger Penrose
Department of Mathematics, King's College, London, England
(Received 14 December 1962)

This note outlines a new technique for studying asymptotic questions in (special or) general relativity whereby several new results are obtained. The questions dealt with here are the following: (1) a geometrical definition of asymptotic flatness, (2) covariant definitions of incoming and outgoing gravitational (and other) radiation fields, (3) simple deduction of detailed asymptotic behavior of the Riemann tensor (and other fields)—the "peeling off" property,[1-3] (4) definitions of total energy-momentum and its loss by radiation, with conservation laws, (5) unification of finite and asymptotic versions of the characteristic initial value problem,[2-7] and (6) geometrical derivation of the Bondi-Metzner-Sachs asymptotic symmetry group.[2,4,8] A longer term aim of this approach is for a covariant S-matrix theory incorporating gravitation.

The basic idea is as follows. Asymptotic questions are those relating to the "neighborhood of infinity." From the point of view of the metric structure of space-time, however, there is no such thing as a point at infinity, since such a point would be an infinite distance from its neighbors. But if we think only in terms of conformal structure of space-time (only ratios of neighboring infinitesimal distances are to have significance), then infinity can be treated as though it were simply an ordinary three-dimensional boundary $\mathscr{I}$ to a "finite" four-dimensional conformal region $\mathfrak{M}$. In fact, we may envisage a new "unphysical" metric $g_{\mu\nu}$ assigned (but perhaps only locally) to space-time, which is conformal to the original physical metric $\tilde{g}_{\mu\nu}$ with

$$g_{\mu\nu} = \Omega^2 \tilde{g}_{\mu\nu},$$

and according to which "infinity" is now finite and in most places regular. The boundary $\mathscr{I}$ of $\mathfrak{M}$ is given by $\Omega = 0$, with $\Omega_{;\mu} \neq 0$. ("Infinity" is given finite coordinate values, so $\tilde{g}_{\mu\nu}$ becomes infinite there.)

All covariant derivatives used here will be carried out according to the unphysical $g_{\mu\nu}$ metric so that properties of $\mathscr{I}$ and its neighborhood in $\mathfrak{M}$ may be studied. Any such property which is conformally invariant will then be a physically meaningful asymptotic property in the original space-time $\tilde{\mathfrak{M}}$. The basic reason for success with this approach is the conformal invariance of the zero rest-mass equations for each spin.[9,10] In particu-

lar, the spin-2 equation $K_{\mu\nu[\rho\sigma;\tau]}=0$ is preserved with

$$\tilde{K}^{\mu\nu}{}_{\rho\sigma}=\Omega^{3}K^{\mu\nu}{}_{\rho\sigma}, \tag{1}$$

$K_{\mu\nu\rho\sigma}$ having the symmetries and trace-free properties of an empty-space Riemann tensor (or Weyl tensor).

The physical curvature quantities are obtainable from the unphysical ones by[11]

$$\tilde{S}^{\mu}{}_{\nu}=\Omega^{2}S^{\mu}{}_{\nu}+2\Omega\Omega^{;\mu}{}_{;\nu}-\Omega^{;\rho}\Omega_{;\rho}\delta^{\mu}{}_{\nu}, \tag{2}$$

$$\tilde{C}^{\mu\nu}{}_{\rho\sigma}=\Omega^{2}C^{\mu\nu}{}_{\rho\sigma}, \tag{3}$$

where $S_{\mu\nu}=\frac{1}{6}Rg_{\mu\nu}-R_{\mu\nu}$, and $C_{\mu\nu\rho\sigma}$ is Weyl's tensor, comprising exactly the curvature information not contained in $S_{\mu\nu}$.

Assume there is no λ term in Einstein's equations, and that all fields except gravitation, electromagnetism, and neutrinos are restricted to a bounded region of space. Then $\tilde{R}=0$ in the neighborhood of $\mathcal{I}$, whence (2) implies $\Omega^{;\rho}\Omega_{;\rho}=0$ on $\mathcal{I}$. Thus, $\mathcal{I}$ is a null hypersurface. Consider, in particular, Minkowski space.[10] Here $\mathcal{I}$ can be separated into five distinguishable disjoint parts, namely, three points I^{-}, I^{0}, and I^{+} representing, respectively, the past, spatial, and future infinities, and two null hypersurfaces $\mathcal{I}^{-}$ and $\mathcal{I}^{+}$ representing the past and future null infinities. Each of $\mathcal{I}^{-}$ and $\mathcal{I}^{+}$ has the topology of a three-dimensional cylinder ($S^{2}\times E^{1}$), bounded by I^{-} and I^{0}, and I^{0} and I^{+}, respectively, at its "past" and "future" (see Fig. 1). Furthermore, any null geodesic in $\mathfrak{M}$,

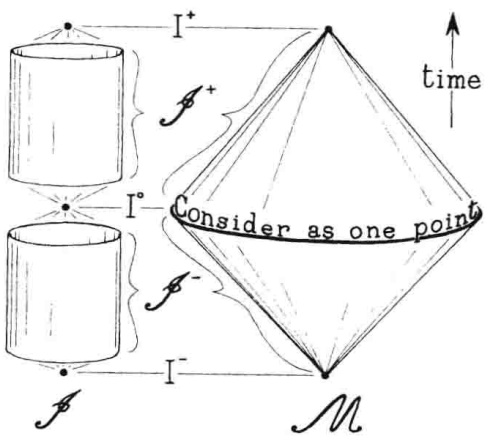


FIG. 1. Conformal structure of infinity.

not on $\mathcal{I}$, originates at a point of $\mathcal{I}^{-}$ and terminates on $\mathcal{I}^{+}$. An examination of Schwarzschild's solution and of the more general radiative metrics[2,4,5] suggests as a suitable global definition of asymptotic flatness[12] for $\tilde{\mathfrak{M}}$, the fact that $\mathfrak{M}$ exists, with the structure as defined above, which is regular everywhere (say C^{5}) up to and including its boundary $\mathcal{I}$ except at I^{-}, I^{0}, and I^{+}.

Many metrics are consistent with a given such conformal structure for $\mathfrak{M}$, and a choice for which $\Omega_{;\mu}=-n_{\mu}$ with $n_{\mu}n^{\mu}=0$ is always locally possible. For simplicity, such a choice will be made here together with a physically reasonable (but probably unnecessary) assumption that $\tilde{R}^{\mu}{}_{\nu}=O(\Omega^{4})$ in the neighborhood of $\mathcal{I}^{+}$ (say). Then (2) implies $\Omega_{;\mu;\nu}=0$ on $\mathcal{I}^{+}$, whence the shear and divergence[1,3] of $\mathcal{I}^{+}$ vanish, this vanishing of shear being a conformally invariant property. Differentiating, $\Omega_{;\mu;\nu;\rho}=\frac{1}{2}S_{\mu\nu}n_{\rho}$, $\Omega_{;\mu;\nu;\rho;\sigma}=S_{\mu\nu;(\rho}n_{\sigma)}$, $n^{\mu}S_{\mu\nu}=0$, $n^{\mu}S_{\mu\nu;\rho}=0$ on $\mathcal{I}^{+}$, whence Ricci identities give $C_{\mu\nu\rho\sigma}n^{\sigma}=0$, $C_{\mu\nu\rho\sigma}{}^{;\sigma}=0$ on $\mathcal{I}^{+}$. This and the topology of $\mathcal{I}$ imply $C_{\mu\nu\rho\sigma}=0$ on $\mathcal{I}$. Hence the gravitational spin-2 field $K_{\mu\nu\rho\sigma}$ [see (1), (3)] can be defined continuously throughout $\mathfrak{M}$ by $\Omega K_{\mu\nu\rho\sigma}=C_{\mu\nu\rho\sigma}$ with $-K_{\mu\nu\rho\sigma}n_{\tau}=C_{\mu\nu\rho\sigma;\tau}$ on $\mathcal{I}$. Completing a null tetrad with m_{μ} and $\bar{m}^{\mu}$ complex, l_{μ} real satisfying $m_{\mu}m^{\mu}=l_{\mu}l^{\mu}=l_{\mu}m^{\mu}=n_{\mu}m^{\mu}=1-l_{\mu}n^{\mu}=1+m_{\mu}\bar{m}^{\mu}=0$, we can define the outgoing gravitational radiation field as the complex tetrad component[3] $\Psi_{4}=K_{\mu\nu\rho\sigma}n^{\mu}\times m^{\nu}n^{\rho}m^{\sigma}$ on $\mathcal{I}^{+}$. This is uniquely determined, except for scaling, by the conformal geometry of $\mathfrak{M}$. The incoming field is defined similarly on $\mathcal{I}^{-}$. (The electromagnetic radiation field is $F_{\mu\nu}n^{\mu}m^{\nu}$; the neutrino radiation field is also simply defined.)

Let l be a null geodesic meeting $\mathcal{I}^{+}$ at P. Then, in the neighborhood of P, $g_{\mu\nu}$ may be chosen so that $-\Omega$ and $\tilde{r}=\Omega^{-1}$ are affine parameters on l according to $g_{\mu\nu}$ and $\tilde{g}_{\mu\nu}$, respectively. Choosing l^{μ} tangent to l, and m^{μ} suitably, the "tetrads" $(l^{\mu}, m^{\mu}, n^{\mu})$, $(\tilde{l}^{\mu}=\tilde{r}^{2}l^{\mu}, \tilde{m}^{\mu}=\tilde{r}m^{\mu}, \tilde{n}^{\mu}=n^{\mu})$ are transported parallelly along l according to $g_{\mu\nu}$ and $\tilde{g}_{\mu\nu}$, respectively. It follows at once from continuity of $K_{\mu\nu\rho\sigma}$ that $\tilde{\Psi}_{4}=O(\tilde{r}^{-1})$ and that the remaining complex tetrad components $\tilde{\Psi}_{i}$ of $\tilde{K}_{\mu\nu\rho\sigma}$ $(i=0,\cdots,3)$, appropriately ordered, are $O(\tilde{r}^{-5-i})$. This is the "peeling off" property of the Riemann tensor.[1-3] A similar result holds for the electromagnetic and neutrino fields.

Further specialize $g_{\mu\nu}$ so that $R=12$ on $\mathcal{I}^{+}$. Consider any spanning hypersurface $\mathcal{S}$ which meets $\mathcal{I}$ in a spherelike region S, containing one point of each generator of $\mathcal{I}^{+}$. (This implies $\mathcal{S}$ is asymptotically null in $\tilde{\mathfrak{M}}$.) Then the total energy-momentum intercepted by $\mathcal{S}$ is (compare references 2, 4,

and 5)

$$P(W) = (8\pi)^{-1/2} \int (\sigma N - \Psi_2) W dS,$$

where σ is the shear[1,3] of the null hypersurface meeting $\mathcal{I}$ in S and whose tangent vector l_μ serves to define[3] m_μ, Ψ_2 and where

$$N = -\tfrac{1}{2} R_{\mu\nu} \bar{m}^\mu \bar{m}^\mu$$

is essentially Bondi's "news function." It is also the derivative in the l_μ direction at $\mathcal{I}^+$ of the shear of $\Omega = \text{const}$ and it has a conformally invariant interpretation. W is a real weighting factor satisfying $1 - W^2 = W_{;\mu;\nu} m^\mu \bar{m}^\nu W - |W_{;\mu} m^\mu|^2$ and has four degrees of freedom (corresponding to different possible "time axes"), one choice ("energy") being $W = 1$. The others are generated by the different permissible choices of metric $g_{\mu\nu}$, and with appropriate interpretations, $P(W)$ behaves as a 4-vector. Taking the difference between $P(W)$ and the corresponding value for a "later" hypersurface leads to a conservation law with the definition of gravitational energy-momentum flux across $\mathcal{I}^+$ as $N\bar{N}$. (The electromagnetic energy-momentum flux is $|F_{\mu\nu} m^\mu n^\nu|^2$, and correspondingly for the neutrino field.)

Initial data for gravitation may be specified on $\mathcal{S}$, and N specified[2,4,5] on the part of $\mathcal{I}^+$ "below" S. Essentially equivalent, however, is to use $\Psi_4 = N_{;\mu} n^\mu$ on $\mathcal{I}^+$ and, if $\mathcal{S}$ is null, to use Ψ_0 on $\mathcal{S}$ (and complete the data by giving certain quantities on S). The exact analogy between Ψ_4 on $\mathcal{I}^+$ and Ψ_0 on $\mathcal{S}$ leads to a unification of the finite[6,7] and asymptotic[2,4,5] versions of the characteristic initial value problem.

The Bondi-Metzner-Sachs[2,4,8] asymptotic symmetry group for general relativity can be interpreted as a group of conformal self-transformations of the three-dimensional manifold $\mathcal{I}^+$ (or $\mathcal{I}^-$). Conformal transformations always preserve finite angles, but there are also the null angles on $\mathcal{I}^+$ between tangent vectors of which n^μ is a linear combination. Parallel transport establishes an equivalence relation between null angles which turns out to be independent of the choice of $g_{\mu\nu}$. The required group now consists of the self-transformations of $\mathcal{I}^+$ (or $\mathcal{I}^-$) which preserve both angles and null angles (and do not reverse time-sense). If any of I^-, I^0, and I^+ were nonsingular, then the inhomogeneous Lorentz subgroup could be singled out, but this is not generally the case.[13]

A full account of these results will be published elsewhere.

---

*Work partly supported by the Aeronautical Research Laboratory, Office of Air Research, through the European Office of the U. S. Air Force Aerospace Research.

[1]R. K. Sachs, Proc. Roy. Soc. (London) A264, 309 (1961).

[2]R. K. Sachs, Proc. Roy. Soc. (London) A270, 103 (1962).

[3]E. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962).

[4]H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, Proc. Roy. Soc. (London) A269, 21 (1962).

[5]E. Newman and T. W. J. Unti, J. Math. Phys. 3, 896 (1962).

[6]R. K. Sachs, J. Math. Phys. 3, 908 (1962).

[7]R. Penrose, Null Hypersurface Initial Data for Classical Fields of Arbitrary Spin and for General Relativity, 1961 (to be published).

[8]R. K. Sachs, Asymptotic Symmetries in Gravitational Theory, 1962 (to be published).

[9]E. Cunningham, Proc. London Math. Soc. 8, 77 (1910); J. A. Schouten, Rev. Modern Phys. 21, 421 (1949).

[10]R. Penrose, Report to the Warsaw Conference, 1962 (unpublished).

[11]See, for example, J. A. Schouten, Ricci-Calculus (Springer-Verlag, Berlin, Germany, 1954).

[12]For other approaches, see references 2, 4, 5, and 8, and, for example, A. Trautman, Bull Acad. Polon. Sci. III 6, 403, 407 (1958); P. G. Bergmann, I. Robinson, and E. Schucking, Phys. Rev. 126, 1227 (1962); R. Arnowitt, S. Deser, and C. W. Misner, Phys. Rev. 122, 997 (1961).

[13]Compare also P. G. Bergmann, Phys. Rev. 124, 274 (1961).

17
THE LIGHT CONE AT INFINITY

Relativistic Theories of Gravitation. Ed. L. Infeld 1964,
pp. 369–373, Pergamon Press, Oxford

This was actually the first article I wrote on conformal infinity, and its scope is not nearly so extensive as Chapter 16, which was written somewhat later. The 2-dimensional case described in Figure 1 resembles what is sometimes referred to as a 'Penrose diagram'. It is not, however a 'strict conformal diagram' (or 'Carter diagram') in the specific sense described on p. 725 of my book, *The Road to Reality* (2004; Jonathan Cape, London).

442 The light cone at infinity. *Relativistic Theories of Gravitation*. Ed. L. Infeld 1964, pp. 369–373

THE LIGHT CONE AT INFINITY

R. PENROSE

King's College, London

Abstract—From the point of view of the conformal structure of space-time, "points at infinity" can be treated on the same basis as finite points. Minkowski space can be completed to a highly symmetrical conformal manifold by the addition of a null cone at infinity—the "absolute cone". Owing to their conformal invariance, zero rest-mass fields can be studied on the whole of this conformal manifold. The behaviour of the fields on the absolute cone can be used to define the radiation fields. General relativity can be treated from this point of view if the gravitational field is represented by an expression originally equal to the Weyl tensor but which transforms slightly differently under conformal transformation.

QUESTIONS concerning radiation, or asymptotic flatness of space-time, involve statements about events in the "neighbourhood of infinity". It would appear, therefore, that some deeper understanding of the mathematical nature of this "infinity" might be of great conceptual value to physics. I wish to describe here an idea which seems to have been partially suggested by a number of people, and to me particularly by E. Schücking. The essential construction can apparently be traced back to Möbius [1].(1)

The idea is that if space-time is considered from the point of view of its conformal structure only, points at infinity can be treated on the same basis as finite points. It should be recalled that the concepts of angle, of the light cone at any event and a null geodesic, are conformally invariant and therefore pertinent to the conformal geometry of space-time. The concepts of infinitesimal distance and of space-like or time-like geodesics are not.

I shall be concerned here primarily with Minkowski space and its completion to a compact conformal manifold. The construction is analogous to the completion of a Euclidean plane to a projective plane by the addition of a "line at infinity", or alternatively, to its completion to an inversive plane by the addition of a "point at infinity". Some considerations applicable to curved space-time will be given at the end.

(1) Another article of much greater immediate relevance has recently been brought to my attention. This is: Hans Rudberg, The compactification of a Lorentz space and some remarks on the foundation of the theory of conformal relativity; dissertation (1957), University of Uppsala; Physics Abstracts No 30 Vol. 61 (1958). In it, the geometry of the conformally completed Minkowski space is discussed in some considerable detail.

The light cone at infinity. *Relativistic Theories of Gravitation*. Ed. L. Infeld 1964, pp. 369–373

Let x^μ be the position vector of a general event in Minkowski space-time relative to a given origin O. Then the transformation to new Minkowskian coordinates $\hat{x}^\mu$ given by

$$\hat{x}^\mu = \frac{x^\mu}{x_a x^a}, \qquad x^\mu = \frac{\hat{x}^\mu}{\hat{x}_a \hat{x}^a}, \tag{1}$$

is *conformal* ("inversion with respect to O"). Observe that the whole null cone of O is transformed to infinity in the $\hat{x}^\mu$ system and that infinity in the x^μ system becomes the null cone of the origin $\hat{O}$ of the $\hat{x}^\mu$ system. ("Space-like" or "time-like" infinity become $\hat{O}$ itself but "null" infinity becomes spread out over the null cone of $\hat{O}$.) Thus, from the conformal point of view "infinity" must be a *null cone*. The two systems x^μ, $\hat{x}^\mu$ related by (1) may be regarded as two coordinate systems, each of which covers part of a conformally flat and *compact* manifold $\mathcal{M}$. Together they do not cover quite the whole of $\mathcal{M}$ since the points at infinity on the null cone of O are excluded in both coordinate systems, but these points can easily be covered by choosing a third coordinate system which is related to $x^\mu - a^\mu (a_a a^a > 0)$ in the same way that $\hat{x}^\mu$ is related to x^μ in (1).

The geometry of $\mathcal{M}$ is then briefly as follows. $\mathcal{M}$ contains ∞^4 points and ∞^5 null (straight) lines. The ∞^2 null lines through each point generate the null cone of that point. These null cones are all closed, the null lines being topologically circles. Each cone has just one vertex. $\mathcal{M}$ admits a transitive ∞^{15} group of motions, so that all its points are on the same footing. Thus all these null cones are also on the same footing. If any one of these cones is chosen, it may be regarded as an *absolute cone* (cone at infinity) for a Minkowski metric structure consistent with the given conformal structure. The metrical concepts can all be defined in relation to this absolute cone. Thus, "parallel" null lines are null lines which meet the same generator of the absolute cone. If they meet at the same point of the absolute cone, they are not only "parallel" but they lie in the same "null hyperplane". Thus, a "null hyperplane" is simply a null cone whose vertex lies on the absolute cone (other than at its vertex). $\mathcal{M}$ also contains ∞^9 *space-like circles* and ∞^9 *time-like circles*—a time-like circle being, in general, the world-line of a uniformly accelerating particle (together with the other branch of the "hyperbola"). A space-like or time-like "straight line" is simply a space-like or time-like circle which passes through the vertex of the absolute cone. (Note the characteristic fact that on the other hand "null straight lines" will *not* pass through the vertex of the absolute cone, if "finite".) A limiting case of a space-like or time-like circle is a pair of intersecting null lines.

To picture $\mathcal{M}$ consider first the two-dimensional case. Imagine the whole of two-dimensional Minkowski space-time to be mapped continuously onto the interior of a square, with the null lines parallel to the sides (see Fig. 1).

 The light cone at infinity. *Relativistic Theories of Gravitation*. Ed. L. Infeld 1964, pp. 369–373

Then infinity is represented by the sides of the square and to complete the picture, opposite sides must be identified preserving sense. The resultant compact manifold is topologically a torus. Next, consider the three-dimen-

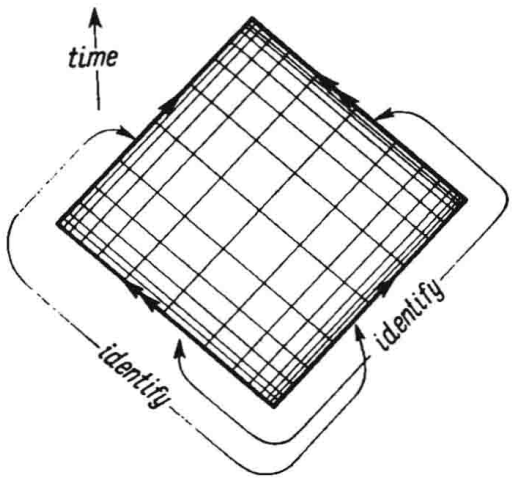


FIG. 1.

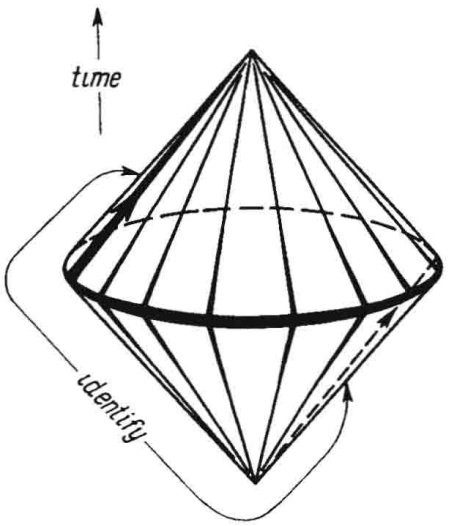


FIG. 2.

sional case (two space and one time dimension). This time we map the space-time continuously onto the interior of a region bounded by two portions of cones joined base to base (see Fig. 2). Each generator of the top cone is to be identified with the opposite generator of the bottom cone, preserving future sense. It follows that both the top vertex and the bottom vertex must be identified with the whole of the "equator"—which must therefore be considered as a point. The resultant compact manifold is non-orientable and has the topology of a three-dimensional analogue of Klein's bottle. The four-dimensional case is very similar, except that in this case the manifold turns out to be orientable again and has the topology $S^1 \times S^3$.

Note the fact that the removal of any null cone from $\mathcal{M}$ leaves a (simply-) connected set of points. This is most easily seen if the cone is chosen to be the absolute cone. Thus the three regions "past", "future", "elsewhere" into which a null cone divides normal Minkowski space are *connected* to each other in $\mathcal{M}$. There is, thus, no invariant distinction between a space-like or a time-like separation for two general points in $\mathcal{M}$.

The possibility of applying these ideas to physics rests on the fact that any zero rest-mass free field can be regarded (with a suitable interpretation) as being conformally invariant. Thus, under the conformal transformation

$$\tilde{g}_{\mu\nu} = \lambda^2 g_{\mu\nu}, \tag{2}$$

λ being a function of position, the source-free Maxwell equations, in particular, are preserved if we put

$$\tilde{F}_{\mu\nu} = F_{\mu\nu}.$$

For the case of a source-free spin-two field we have a tensor $K_{\mu\nu\rho\sigma}$ with the symmetries and trace-free conditions of an empty-space Riemann tensor (or Weyl tensor) satisfying

$$K_{\mu\nu[\rho\sigma;\tau]} = 0. \tag{3}$$

This is preserved under the conformal transformation (2) if we put

$$\tilde{K}_{\mu\nu\rho\sigma} = \lambda K_{\mu\nu\rho\sigma} \tag{4}$$

so that the field must be regarded as a suitable kind of density. The case of general spin is most easily treated in the two-component spinor formalism. For any spin $s > 0$ we can describe the field by a symmetric spinor $\theta_{AB\cdots E}$ with $2s$ indices satisfying

$$\theta_{AB\cdots E;\mu}\sigma^{\mu A\dot{P}} = 0. \tag{5}$$

Then, (5) is preserved under the conformal transformation (2) (accompanied by $\tilde{\sigma}_{\mu}{}^{A\dot{P}} = \lambda\sigma_{\mu}{}^{A\dot{P}}$) provided

$$\theta_{AB\cdots E} = \lambda^{-s-1}\theta_{AB\cdots E}.$$

Spin-zero, of course, requires special consideration.

Using this conformal invariance it is possible to give a meaning to the concept of a zero rest-mass field defined over the *whole* of $\mathscr{M}$. The condition that the field be *finite* on the absolute cone is a simple statement of an asymptotic condition that is reasonable to impose on the field. Also, it is known that initial data for such a field can be given on any null cone:—one complex number per point of the cone [2]. Hence, we can use the *absolute* cone on which to specify initial data. This initial data then simply measures the strength of the radiation field.

If interactions are to be present, we will not expect the incoming field to match the outgoing field. Therefore, the identification of the past infinity with the future infinity that was done to define the manifold $\mathscr{M}$ seems inadvisable in this case. If the identification is not carried out, we have two absolute cones which bound the Minkowski space, one in the past and one in the future. The comparison between the data on the past cone with that on the future cone determines a kind of S-matrix theory.

In general relativity, the Weyl tensor $C_{\mu\nu\rho\sigma}$ in empty space satisfies the spin-two zero rest-mass free field equations (3), so we may set

$$K_{\mu\nu\rho\sigma} = C_{\mu\nu\rho\sigma}.$$

However, under transformation (2) we have

$$\tilde{C}_{\mu\nu\rho\sigma} = \lambda^2 C_{\mu\nu\rho\sigma}$$

which by comparison with (4) gives

$$\tilde{C}_{\mu\nu\rho\sigma} = \lambda\tilde{K}_{\mu\nu\rho\sigma} \tag{6}$$

so that the Weyl tensor transforms differently from a zero rest-mass spin-two field under conformal transformation. (This is perhaps not surprising,

since Ricci tensor, i.e. "source" for the Weyl field, may be introduced by a conformal transformation). Thus, we must choose for our expression representing the gravitational field, not exactly the Weyl tensor, but the $K_{\mu\nu\rho\sigma}$ which equals the Weyl tensor in the original metric space and which transforms according to (4) under conformal transformation. A reasonable asymptotic condition on the field—a condition of *asymptotic flatness*—can then be stated as the fact that $K_{\mu\nu\rho\sigma}$ must be finite (and suitably regular) on the absolute cone(s). The gravitational field is a (self-)interacting field, so it seems unreasonable to carry out an identification of the infinite past with the infinite future in general relativity—we have two cones, one in the past and one in the future. The strength of the incoming, or outgoing, radiation field is then measured by the initial data for $K_{\mu\nu\rho\sigma}$ on the past, or future, absolute cone.

From (6) it follows that the Weyl tensor *vanishes* on the absolute cones. This is very fortunate since it implies that the conformal structure of infinity is the same in such asymptotically flat curved spaces as in Minkowski space—with the one important difference that the (double) vertex of each absolute cone becomes singular and so is best removed. The absolute cones are then perhaps better thought of as *cylinders* ($S^2 \times E^1$). The possibility of an S-matrix theory for gravitation suggests itself.

REFERENCES

[1] A. F. Möbius, Collected Works, vol. II.

[2] R. Penrose, Null Hypersurface Initial Data for Classical Fields of Arbitrary Spin and for General Relativity, preprint 1961.

18
CONFORMAL TREATMENT OF INFINITY

Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

Again, this is a discussion of conformal infinity, and it goes further than Chapter 16 in a number of respects. The energy-momentum conservation law illustrated in Figure 5 goes further than the situations that had been described previously in the literature. Here, the two cross-sections of I^+ that are depicted need not be time-translations of one another. This generalized Bondi–Sachs conservation law is given here without proof (a proof is provided in R. Penrose & W. Rindler, *Spinors and Space-Time*, Volume 2; 1982; Cambridge University Press, pp. 424–7). Noteworthy is the use of conformal infinity in *cosmological* situations here, where allowance is made for the presence of a cosmological constant—and also the big bang is represented as a conformal boundary. For the first presentation of these ideas, see Chapter 33, in Volume 2.

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

Lecture I

Let $\tilde{\mathscr{M}}$ denote physical space-time with metric $d\tilde{s}$. The idea is to construct another "unphysical" manifold $\mathscr{M}$ with a boundary $\mathscr{I}$ and metric ds, such that $\tilde{\mathscr{M}}$ is conformal to the interior of $\mathscr{M}$ with $ds = \Omega\, d\tilde{s}$, and so that the "infinity" of $\tilde{\mathscr{M}}$ is represented by the "finite" hypersurface $\mathscr{I}$ (see Fig. 1). This last property is expressed by the condition that $\Omega = 0$ on $\mathscr{I}$; that is to say, the metric at $\mathscr{I}$ is stretched by an infinite factor in the passage from $\mathscr{M}$ to $\tilde{\mathscr{M}}$ so $\mathscr{I}$ gets mapped to infinity. *Asymptotic* properties of $\tilde{\mathscr{M}}$ and of fields in $\tilde{\mathscr{M}}$ can now be investigated by studying $\mathscr{I}$, and the *local* behaviour of the fields at $\mathscr{I}$—provided that all the relevant concepts can be put into a conformally invariant form.

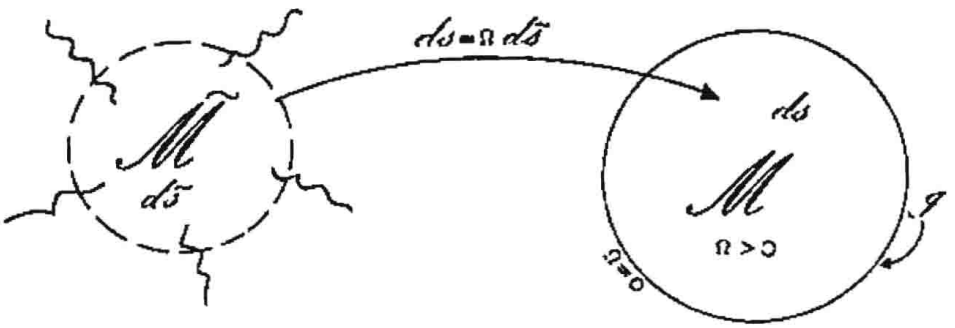


Fig. 1. The infinite physical space-time $\tilde{\mathscr{M}}$ is mapped into an unphysical "finite" conformally equivalent manifold $\mathscr{M}$, with boundary $\mathscr{I}$ corresponding to the "infinity" of $\tilde{\mathscr{M}}$.

The utility of this idea rests on the fact that the zero rest-mass free-field equations for each spin value are conformally invariant if interpreted suitably. For example, for spin zero, if the wave equation is written as

$$\left\{\nabla_\mu \nabla^\mu + \frac{R}{6}\right\}\phi = 0$$

where R is the scalar curvature and $\nabla\mu$ denotes covariant derivative—both according to the metric $g_{\mu\nu}$ of $\mathscr{M}$, then

$$\left\{\tilde{\nabla}_\mu \tilde{\nabla}^\mu + \frac{\tilde{R}}{6}\right\}\tilde{\phi} = 0$$

where $\tilde{\nabla}_\mu$, $\tilde{R}$ refer to the metric $\tilde{g}_{\mu\nu} = \Omega^{-2} g_{\mu\nu}$ of $\mathscr{M}$ and where

$$\tilde{\phi} = \Omega\phi.$$

For spin 1 we have Maxwell's free-field equations

$$\nabla^\mu F_{\mu\nu} = 0, \quad \nabla_{[\lambda} F_{\mu\nu]} = 0$$

with $F_{\mu\nu} = F_{[\mu\nu]}$. The tilde version of this holds if we put

$$\tilde{F}_{\mu\nu} = F_{\mu\nu}.$$

For spin 2, we can use a tensor $K_{\mu\nu\rho\sigma}$ with the symmetries

$$K_{\mu\nu\rho\sigma} = K_{[\rho\sigma][\mu\nu]}, \quad K_{\mu[\nu\rho\sigma]} = 0, \quad K^\mu{}_{\nu\mu\sigma} = 0$$

satisfying

$$\nabla_{[\gamma} K_{\mu\nu]\rho\sigma} = 0$$

or equivalently $\nabla^\mu K_{\mu\nu\rho\sigma} = 0$. Conformal invariance is achieved if

$$\tilde{K}_{\mu\nu\rho\sigma} = \Omega^{-1} K_{\mu\nu\rho\sigma}.$$

The result for general spin is most readily expressed (and proved) using the two-component spinor formalism. A spin s field (s > 0) can be represented by a totally symmetric spinor $\phi_{AB\ldots K}$ with 2s indices satisfying

$$\nabla^{A\dot{P}} \phi_{AB\ldots K} = 0$$

as its free-field equation. For conformal invariance

$$\tilde{\phi}_{AB\ldots K} = \Omega^{s+1} \phi_{AB\ldots K}.$$

The relationship between $\phi_{A\ldots K}$ and the corresponding tensor quantity is given, for spins 1 and 2, by

$$F_{\mu\nu} = \{\phi_{AB}\epsilon_{\dot{C}\dot{D}} + \epsilon_{AB}\bar{\phi}_{\dot{C}\dot{D}}\}\sigma_\mu{}^{A\dot{C}}\sigma_\nu{}^{B\dot{D}},$$

$$K_{\mu\nu\rho\tau} = \{\phi_{ABCD}\epsilon_{\dot{E}\dot{P}}\epsilon_{\dot{G}\dot{H}} + \epsilon_{AB}\epsilon_{CD}\bar{\phi}_{\dot{E}\dot{F}\dot{G}\dot{H}}\}\sigma_\mu{}^{A\dot{E}}\sigma_\nu{}^{B\dot{F}}\sigma_\rho{}^{C\dot{G}}\sigma_\tau{}^{D\dot{H}}.$$

The proof of conformal invariance is easily established from the general spinor formula

$$\begin{aligned}\tilde{\nabla}_{X\dot{Y}}\tilde{\xi}_{A\ldots K\dot{P}\ldots\dot{S}} = \Omega^{m+1}\nabla_{X\dot{Y}}\xi_{A\ldots K\dot{P}\ldots\dot{S}} + \Omega^m\{(m - \tfrac{1}{2}r)\,\xi_{A\ldots K\dot{P}\ldots\dot{S}}\nabla_{X\dot{Y}}\Omega \\ + \xi_{X\ldots K\dot{P}\ldots\dot{S}}\,\nabla_{A\dot{Y}}\,\Omega + \ldots + \xi_{A\ldots X\dot{P}\ldots\dot{S}}\,\nabla_{K\dot{Y}}\,\Omega + \xi_{A\ldots K\dot{Y}\ldots\dot{S}}\,\nabla_{X\dot{P}}\Omega \\ + \ldots + \xi_{A\ldots K\dot{P}\ldots\dot{Y}}\nabla_{X\dot{S}}\Omega\}\end{aligned}$$

where $\xi_{A\ldots K\dot{P}\ldots\dot{S}}$ has r indices altogether and

$$\tilde{\xi}_{A\ldots K\dot{P}\ldots\dot{S}} = \Omega^m\xi_{A\ldots K\dot{P}\ldots\dot{S}}.$$

The conformal invariance in the tensor case then follows from the spinor result. The quantities $\sigma_\mu{}^{A\dot{B}}$ (Hermitian (2 × 2) matrices for each μ) are assumed to transform as

$$\tilde{\sigma}_\mu{}^{A\dot{B}} = \Omega^{-1}\sigma_\mu{}^{A\dot{B}}$$

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

since this preserves the fundamental equation

$$\sigma_\mu{}^{A\dot{B}}\sigma_\nu{}^{C\dot{D}}\epsilon_{AC}\epsilon_{\dot{B}\dot{D}} = g_{\mu\nu}.$$

As a first example of a space-time $\tilde{\mathscr{M}}$ with a finite image $\mathscr{M}$, consider Minkowski space-time. The physical metric can be written, using null coordinates u,v (advanced and retarded times)

$$d\tilde{s}^2 = du\,dv - \tfrac{1}{4}(u - v)^2\{d\theta^2 + \sin^2\theta\, d\phi^2\},$$

$u \geqslant v$. A convenient conformal factor is given by

$$\Omega = (1 + u^2)^{-1/2}(1 + v^2)^{-1/2}.$$

Changing coordinates by

$$u = \tan p, \quad v = \tan q,$$

the unphysical metric takes the form

$$ds^2 = \Omega^2 d\tilde{s}^2 = dp\,dq - \tfrac{1}{4}\sin^2(p - q)\{d\theta^2 + \sin^2\theta\, d\phi^2\},$$

where p and q vary from $-\pi/2$ to $\pi/2$ with $p \geqslant q$. The hypersurface boundary $\mathscr{I}$ of $\mathscr{M}$ is given when $q = -\pi/2$ or $p = \pi/2$, since these values correspond to infinity ($v = -\infty$ or $u = \infty$) for $\tilde{\mathscr{M}}$. Denote the region given by $q = -\pi/2$, $-\pi/2 < p < \pi/2$ by $\mathscr{I}^-$. Similarly let $\mathscr{I}^+$ be given by $p = \pi/2$, $-\pi/2 < q < \pi/2$. The point given by $p = q = -\pi/2$ is denoted by I^-, $p = q = \pi/2$ by I^+ and $p = -q = \pi/2$ by I° (see Fig. 2). The fact that I^-, I^+ and I° are *single points* follows from the fact that $\sin(p - q) = 0$ there. These are coordinate singularities of the type encountered at the origin of polar coordinates. The manifold $\mathscr{M}$ is regular at these points. In fact, the metric ds is that of an Einstein static universe—a cylinder ($S^3 \times E^1$) which represents a three-dimensional spherical space which is constant in time. The manifold $\mathscr{M}$ is only a finite portion of this cylinder (see Fig. 3). The hypersurface $\mathscr{I}^-$ is a future null cone with vertex I^-. This cone is refocused to a point I° which is spacially diametrically opposite to I^-. The future null cone of I° is $\mathscr{I}^+$ which is focused again at I^+. The region lying between $\mathscr{I}^-$ and $\mathscr{I}^+$ is conformal to the entire Minkowski space-time.

The meaning of I^-, $\mathscr{I}^-$, I°, $\mathscr{I}^+$, I^+, can be seen by considering the behaviour of curves in $\mathscr{M}$ corresponding to the straight lines in $\tilde{\mathscr{M}}$. The image of a time-like straight line originates at I^- and terminates at I^+; the image of a space-like straight line originates and terminates at I°; the image of a null straight line originates at a point of $\mathscr{I}^-$ and terminates at a point of $\mathscr{I}^+$. Thus, at I^- represents past infinity; I° represents spatial infinity; I^+ represents future infinity; $\mathscr{I}^-$ represents past *null* infinity; $\mathscr{I}^+$ represents future *null* finity. We expect, therefore, that zero rest-mass fields are to be significant at $\mathscr{I}^-$ and $\mathscr{I}^+$, but that fields of finite rest-mass are important at I^- and I^+ and not at $\mathscr{I}^-$ and $\mathscr{I}^+$.

Consider next the Schwarzschild metric:

$$d\tilde{s}^2 = du^2\left(1 - \frac{2m}{r}\right) + dr\,du - r^2\{d\theta^2 + \sin^2\theta\, d\phi^2\}$$

where one null (retarded) coordinate u has been used. Introducing $ds = \Omega d\bar{s}$ with $\Omega = r^{-1}$ and choosing a new coordinate $l = r^{-1}$ we get

$$ds^2 = du^2(l^2 - 2ml^3) - dl\,du - d\theta^2 - \sin^2\theta\, d\phi^2.$$

This is regular (C^∞) at the hypersurface $\mathscr{I}^+$ defined by $l = 0$, t finite. Similarly by re-expressing the metric in terms of an advanced parameter v in place of the retarded

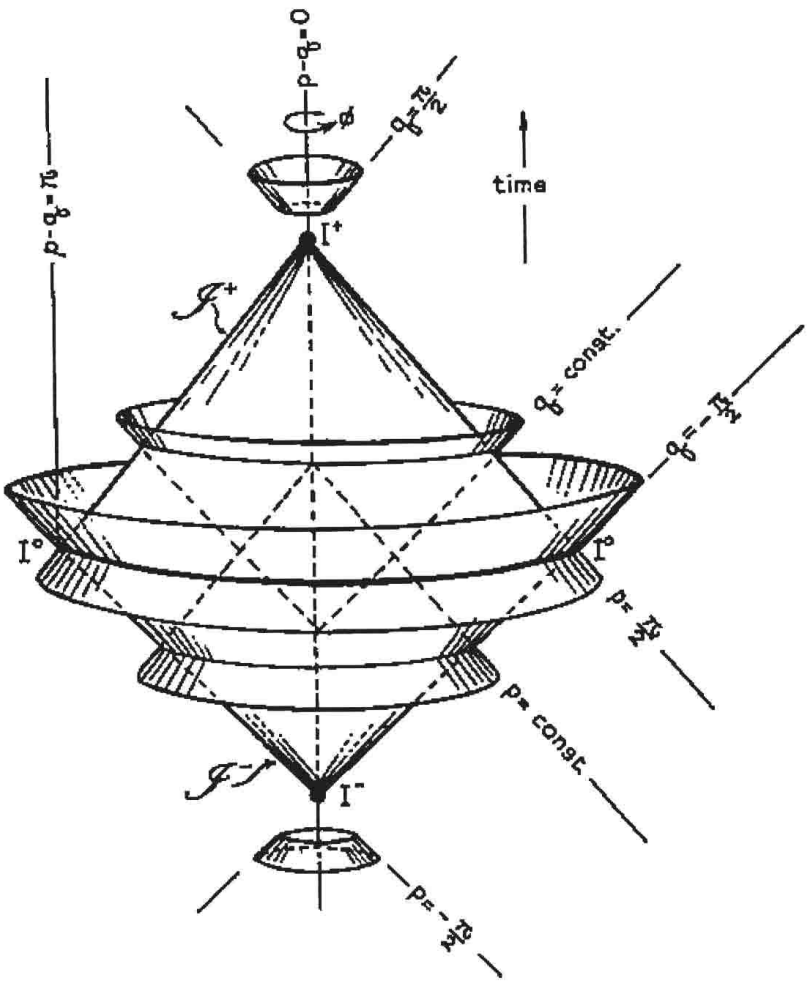


Fig. 2. Coordinates for $\mathscr{M}$ in Minkowski case. The θ coordinate has been suppressed ($\theta = \pi/2$).

parameter u, we get $\mathscr{I}^-$ as regular (C^∞). In each case $\mathscr{I}^+$, $\mathscr{I}^-$ will be topologically a three-dimensional open cylinder ($S^2 \times E^1$). This is precisely the same situation as in the Minkowski case if we exclude the points I^-, I°, I^+. However, if we wish to *include* these points in the manifold $\mathscr{M}$ for the Schwarzschild case, a new situation arises. This is that each of I^-, I°, I^+ must apparently be necessarily a *singular* point for $\mathscr{M}$. That I^- and I^+ are singular is not surprising since all the matter is concentrated at these points when it meets the boundary. However, it also seems that I°

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

must also be singular for any asymptotically flat case where the *total mass* is non-zero—that is to say, in any physically interesting asymptotically flat case except for Minkowski space-time. The fact that these points are singular for $\mathscr{M}$ has importance in relation to the question of asymptotic symmetry groups. This will be discussed briefly in Lecture III.

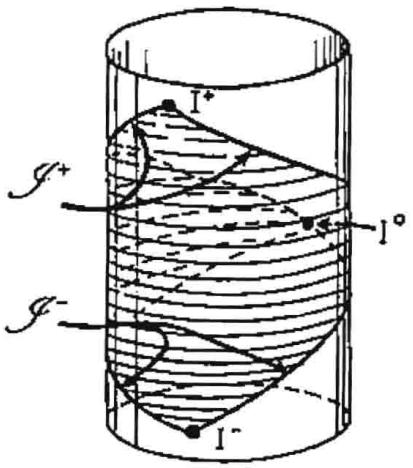


Fig. 3. The cylinder represents the Einstein universe. The two spacial coordinates θ, ϕ have been suppressed; only one space and one time dimension are depicted. The shaded region corresponds to the manifold $\mathscr{M}$.

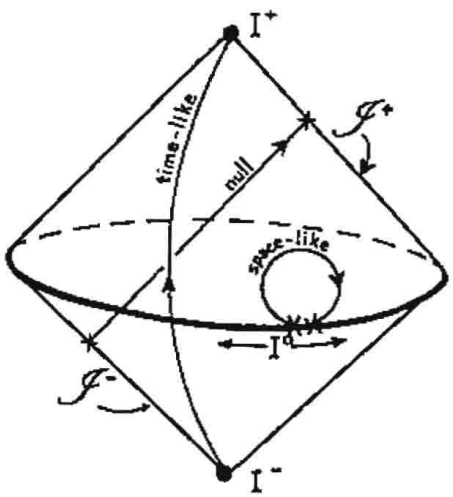


Fig. 4. The images of straight lines in $\mathscr{M}$. (One space dimension is suppressed).

The form of Schwarzschild metric given above is a special case of

$$d\tilde{s}^2 = r^{-2}A\,dr^2 - 2B_i dx^i\,dr + r^2 C_{ij}dx^i\,dx^j$$

$(i, j = 1, 2, 3)$, where if $r^{-1} = x^0$, then A, B_i, C_{ij} are to be sufficiently differentiable functions of x^μ $(\mu = 0, \ldots, 3)$—particularly at the hypersurface $\mathscr{I}$. Then if the

relevant determinant in A, B_l, C_{lj} does not vanish, this means that the metric $ds = \Omega d\hat{s}$ with $\Omega = r^{-1}$ is suitably regular at $\mathscr{I}$. Note that $\nabla_\mu\Omega \neq 0$ for this metric. This condition is important for certain applications. The above metric includes as special cases many which have been suggested as representing gravitational radiation in asymptotically flat space-times, such as those of Bondi, Robinson–Trautman, Sachs and probably also Newman–Unti. Also, de-Sitter space-time is included and so are the asymptotically de-Sitter cases studied by Demianski which are generalizations of the Robinson–Trautman metrics incorporating a cosmological constant.

Lecture II

The notion of a type of space-time $\tilde{\mathscr{M}}$ under consideration here will be made a little more precise at this point. With the notation given in Lecture I, we require:

(a) $\mathscr{M}$, $g_{\mu\nu}$ are sufficiently differentiable;

(b) Ω is sufficiently differentiable throughout $\mathscr{M}$ with $\Omega = 0$ on $\mathscr{I}$ and $\nabla_\mu\Omega \neq 0$ at $\mathscr{I}$ $(= \partial\mathscr{M})$;

(c) every null geodesic in $\mathscr{M}$ has two end-points on $\mathscr{I}$.

Then $\tilde{\mathscr{M}}$ is called *asymptotically simple*. Asymptotically flat and asymptotically de-Sitter space-times are particular cases, but for many applications, condition (c) ought perhaps to be replaced by some weaker local statement. In particular, in order to qualify to be asymptotically simple, the mass singularity of Schwarzschild's solution would have to be replaced by a sufficiently extended region where $R_{\mu\nu} \neq 0$. For otherwise there are the circular null orbits, for example, at $r = 3m$, which do not escape to $\mathscr{I}^+$. The points I^-, I^0, I^+ are here excluded from the precise definition of $\mathscr{M}$. The reason for condition (c) is to ensure that the *whole* of null infinity has been included in $\mathscr{I}$. One important property of conformal transformations should be borne in mind in relation to condition (c). This is that any null geodesic in $\mathscr{M}$ corresponds to a null geodesic in $\tilde{\mathscr{M}}$. The concept of a *null geodesic* (as opposed to that of a space-like or time-like geodesic) is *conformally invariant*. Hence, the past and future infinities of any null geodesic in $\tilde{\mathscr{M}}$ are represented as points of $\mathscr{I}$.

In Lecture III some cases will be considered in which condition (b) does not hold. However, under certain circumstances the condition $\nabla_\mu\Omega \neq 0$ can be *deduced* from the other assumptions of asymptotic simplicity. One such circumstance is given when Einstein's empty-space equations hold (for $\tilde{\mathscr{M}}$) in the neighbourhood of $\mathscr{I}$. Another is given whenever a cosmological constant λ is present, provided it is assumed that if $\lambda > 0$ the energy density is non-negative definite and if $\lambda < 0$, the pressure densities are non-negative definite. Under these circumstances it can also be shown that

$$\lambda > 0 \textit{ implies } \mathscr{I} \textit{ is space-like},$$

$$\lambda < 0 \textit{ implies } \mathscr{I} \textit{ is time-like}$$

and furthermore, if $\nabla_\mu\Omega \neq 0$, then

$$\lambda = 0 \textit{ implies } \mathscr{I} \textit{ is null}.$$

The radiation fields of the various zero rest-mass fields in $\tilde{\mathscr{M}}$ can be studied by examining the behaviour of the conformally transformed fields at points of $\mathscr{I}$. In

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

order to do this in the gravitational case, it will be necessary to represent the gravitional field as a spin 2 field of the type considered in Lecture I. We observe that if Einstein's empty-space equations hold—with or without cosmological constant—then Weyl's conformal tensor $\tilde{C}_{\mu\nu\rho\sigma}$ for $\tilde{\mathscr{M}}$ (defined by Dr. Sach's in his lectures) satisfies

$$\tilde{\nabla}_{[\lambda}\tilde{C}_{\mu\nu]\rho\sigma} = 0$$

and possesses the symmetries

$$\tilde{C}_{\mu\nu\rho\sigma} = \tilde{C}_{[\rho\sigma][\mu\nu]}, \quad \tilde{C}_{\mu[\nu\rho\sigma]} = 0, \quad \tilde{C}^{\mu}{}_{\nu\mu\sigma} = 0.$$

The spinor equivalent of $C_{\mu\nu\rho\sigma}$ is Ψ_{ABCD} defined by

$$C_{\mu\nu\rho\tau} = \{\Psi_{ABCD}\epsilon_{\dot{E}\dot{F}}\epsilon_{\dot{G}\dot{H}} + \epsilon_{AB}\epsilon_{CD}\bar{\Psi}_{\dot{E}\dot{F}\dot{G}\dot{H}}\}\sigma_{\mu}{}^{A\dot{E}}\sigma_{\nu}{}^{B\dot{F}}\sigma_{\rho}{}^{C\dot{G}}\sigma_{\tau}{}^{D\dot{H}}$$

and satisfies

$$\tilde{\nabla}^{A\dot{P}}\tilde{\Psi}_{ABCD} = 0$$

in empty space. If matter is present, then (in general) a right-hand side appears to this equation. We can thus identify a spin 2 field ϕ_{ABCD} in $\tilde{\mathscr{M}}$ with this spinor field Ψ_{ABCD}:

$$\phi_{ABCD} = \tilde{\Psi}_{ABCD} \text{ or } \tilde{K}_{\mu\nu\rho\sigma} = \tilde{C}_{\mu\nu\rho\sigma}.$$

However, under conformal transformation we have

$$\tilde{\Psi}_{ABCD} = \Omega^{+2}\Psi_{ABCD}, \quad \tilde{C}_{\mu\nu\rho\sigma} = \Omega^{-2}C_{\mu\nu\rho\sigma}.$$

Hence, comparing this with the factors obtained in lecture I, we get

$$\phi_{ABCD} = \Omega^{-1}\Psi_{ABCD}, \quad K_{\mu\nu\rho\sigma} = \Omega^{-1}C_{\mu\nu\rho\sigma}.$$

The difference between the conformal behaviour of $K_{\mu\rho\nu\sigma}$ and of $C_{\mu\nu\rho\sigma}$ is really not surprising when we consider that, in general, the Ricci tensor of $\mathscr{M}$ will *not* vanish (nor be proportional to $g_{\mu\nu}$) when $\tilde{\mathscr{M}}$ is empty near $\mathscr{I}$. Thus, the free-field zero rest-mass equation will not generally be satisfied by $C_{\mu\nu\rho\sigma}$ for an empty space-time $\tilde{\mathscr{M}}$, whereas this equation *will* be satisfied by $K_{\mu\nu\rho\sigma}$.

If the gravitional radiation field is to be examined by looking at components of $K_{\mu\nu\rho\sigma}$ at points of $\mathscr{I}$, then we must have $K_{\mu\nu\rho\sigma}$ *finite* (and, say, continuous) at $\mathscr{I}$. It is not clear, as yet, that this will generally be the case since this would require $C_{\mu\nu\rho\sigma} = 0$ at $\mathscr{I}$. We know, at this stage, only that $C_{\mu\nu\rho\sigma}$ is suitably regular at $\mathscr{I}$ because of condition (a). However, we also have

$$\nabla^{A\dot{P}}\{\Omega^{-1}\Psi_{ABCD}\} = 0.$$

Hence,

$$n^{A\dot{P}}\Psi_{ABCD} = \Omega\nabla^{A\dot{P}}\Psi_{ABCD} = 0.$$

at $\mathscr{I}$, where $n^{\mu} = \nabla^{\mu}\Omega$ is normal to $\mathscr{I}$ and is assumed to be non-zero. Now, if a cosmological term is present in Einstein's equations, $\mathscr{I}$ is space-like or time-like. In

this case, $n^{A\dot{P}}$ is non-singular and it follows that $\Psi_{ABCD} = 0$ at $\mathscr{I}$ (and hence $C_{\mu\nu\rho\sigma} = 0$ at $\mathscr{I}$). From C^1 differentiability of Ψ_{ABCD} it follows that $\phi_{ABCD} = \Omega^{-1}\Psi_{ABCD}$ is *continuous at* $\mathscr{I}$. On the other hand, if no cosmological constant is present, $\mathscr{I}$ is null. Thus $n^{A\dot{P}}$ is singular and we obtain only

$$\iota^A\Psi_{ABCD} = 0$$

where $n^{A\dot{P}} = \pm\, \iota^A\bar{\iota}^{\dot{P}}$. By differentiating the equation for $\Omega^{-1}\Psi_{ABCD}$ once further we obtain

$$\iota^E\nabla_{E\dot{P}}\Psi_{ABCD} = 0$$

at $\mathscr{I}$, provided that

$$\nabla_{(A}{}^{\dot{C}}\nabla_{B)}{}^{\dot{D}}\Omega = 0$$

at $\mathscr{I}$. This last equation in fact follows from the formula for the trace-free part of the Ricci tensor of $\mathscr{M}$ when Einstein's empty-space equations hold (near $\mathscr{I}$) for $\tilde{\mathscr{M}}$. It can then be shown, using a global argument, that provided the topology of $\mathscr{I}$ is as it is for the Minkowski and Schwarzschild cases, then $\Psi_{ABCD} = 0$ at $\mathscr{I}$ and hence ϕ_{ABCD} is continuous as before.

It is, in fact, possible to show that when $\mathscr{I}$ is null its topology is *necessarily* of the form $\mathscr{I} = \mathscr{I}^- \cup \mathscr{I}^+$, where $\mathscr{I}^- = \mathscr{I}^+ \cong S^2 \times E^1$, as it is for the case of Minkowski space-time (I^-, I^0, I^+ are omitted). The proof is somewhat involved and it rests heavily on condition (c) for asymptotic simplicity.

When the Einstein–Maxwell equations are assumed to hold in the neighbourhood of $\mathscr{I}$, then the above arguments can be carried through for both the gravitational and the electromagnetic fields, if either the cosmological constant is positive, or it is zero and certain pathological possibilities are excluded. The result obtained is that the gravitational ϕ_{ABCD} is continuous at $\mathscr{I}$ and that the electromagnetic ϕ_{AB} is (at least) bounded at $\mathscr{I}$. No doubt these results can be strengthened somewhat.

The importance of establishing the finiteness of ϕ_{ABCD} and ϕ_{AB} at $\mathscr{I}$, lies in the relationship between this fact and the so-called "peeling-off" property (described by Dr. Sachs in his lectures). Effectively this property means that the various complex components of ϕ_{ABCD} and ϕ_{AB} fall off as successive negative powers of an affine (or luminosity) parameter r in retarded or advanced null directions in $\tilde{\mathscr{M}}$. This in turn is related to the asymptotic behaviour of the principal null directions of the fields.

Let the symmetric spinor $\phi_{A\ldots K}$ with $2s$ indices be bounded near $\mathscr{I}$. Let g be a null geodesic in $\mathscr{M}$ meeting $\mathscr{I}$ at a point G. Let ξ_A be a spinor associated with the tangent direction to g and choose it to be parallelly transported along g. Under conformal transformation this is preserved if

$$\tilde{\xi}_A = \Omega^{1/2}\xi_A.$$

Let η_A be chosen also to be parallelly transported along g, with $\xi_A\eta^A = 1$, and so that, at G, the normal to $\mathscr{I}$ is linearly dependent on the two null directions represented by ξ_A, η_A. We can then arrange $\tilde{\xi}_A\tilde{\eta}^A = 1$, with $\tilde{\eta}_A$ parallelly transported

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

along $\bar{g}$ in $\mathscr{M}$ where

$$\bar{\eta}_A = \Omega^{-1/2}\eta_A + O(\Omega^{1/2}\log\Omega).\xi_A.$$

From the finiteness of each component $\phi_{A\ldots CD\ldots K}\,\eta^A\ldots\eta^C\xi^D\ldots\xi^K$ it follows that

$$\tilde{\phi}_{A\ldots CD\ldots K}\bar{\eta}^A\ldots\bar{\eta}^C\xi^D\ldots\bar{\xi}^K = O(\Omega^{k+1})$$

where there are k $\bar{\xi}$'s on the left and where $\tilde{\phi}_{A\ldots K} = \Omega^{s+1}\phi_{A\ldots K}$ as in lecture I. We also have $\Omega^{-1} \sim ar$ where r is an affine (or luminosity) parameter along g and a is some constant. This, then, gives us the "peeling-off" property for the field $\tilde{\phi}_{A\ldots K}$ in

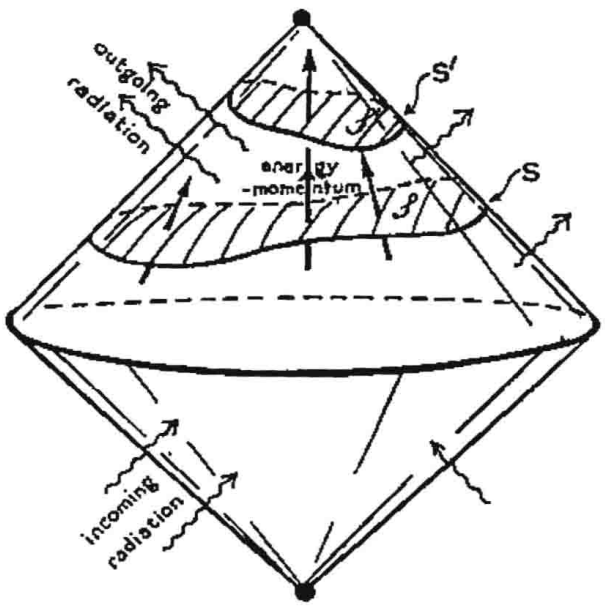


Fig. 5. The energy-momentum intercepted by $\mathscr{S}$ is measured by an integral over S. The corresponding integral over S' is less by an amount equal to the energy-momentum carried away by the radiation which escapes between S and S'.

$\mathscr{M}$. In particular it is of interest to note the *radiation field* component (or "null part") of $\phi_{A\ldots K}$ which is given when $k = 0$. We see from the above that this falls off as r^{-1}.

The other components are also of interest and, in particular, in the gravitational case, the r^{-3} part comes into the definition of total energy-momentum of the system. Let $\mathscr{S}$ be a hypersurface in $\mathscr{M}$ spanning some two-dimensional cross-section S of $\mathscr{I}^+$. That is to say, $S = \dot{\mathscr{S}} = \mathscr{S} \cap \mathscr{I}^+$ where S meets each generator of $\mathscr{I}^+$ in one point and each point of S lies on a generator of $\mathscr{I}^+$. $\mathscr{S}$ will be *asymptotically null* in $\mathscr{M}$. The two-dimensional surface S is to be space-like and will have the topology of a sphere (S^2). Such a surface can always be transformed, by the introduction of a suitable conformal factor, into a metric sphere of (say) unit radius. Suppose that this has been done. Then the *total energy-momentum* intercepted by $\mathscr{S}$

is equal to the following integral over S:

$$P_\mu = \frac{1}{4\pi}\int (\sigma N - \psi_2) W_\mu \, dS$$

where dS is an element of surface area of S,

$$\psi_2 = \phi_{ABCD}\eta^A\eta^B\xi^C\xi^D,$$
$$N = -\tfrac{1}{2}R_{\mu\nu}\sigma^\mu{}_{A\dot{B}}\sigma^\nu{}_{C\dot{D}}\eta^A\bar{\xi}^{\dot{B}}\eta^C\bar{\xi}^{\dot{D}}$$
$$\sigma = \xi^A\bar{\eta}^{\dot{B}}\xi^C\nabla_{A\dot{B}}\xi_C$$

with η^A corresponding to the null direction in $\mathscr{I}^+$, ξ^A corresponding to the other null direction which is orthogonal to dS and normalized such that $\xi_A\eta^A = 1$. Then N is effectively the Bondi–Sachs "news function". The quantity σ is the shear of the null hypersurface which meets $\mathscr{I}^+$ in S (Bondi–Sachs "c"). The quantity $W\mu$ is a real weighting factor which behaves as a 4 vector. We can choose $W_0 = 1$ for the energy component and $W_1 = \cos\theta$, $W_2 = \sin\theta\sin\phi$, $W_3 = \sin\theta\cos\phi$ for the three momentum components, where θ and ϕ are spherical polar coordinates on S.

This expression for energy-momentum is effectively a translation of the results of Bondi and Sachs into the present form. However, it is possible to give an independent derivation of their conservation law using the present formalism. In fact, a slightly more general conservation law can be obtained. Let S' be another (say later) cross-section of $\mathscr{I}^+$. Then the corresponding energy-momentum integral $P'\mu$ can be constructed which represents the total energy-momentum intercepted by $\mathscr{S}'$. The difference $P_\mu - P'_\mu$ can be expressed as an integral over the portion of $\mathscr{I}^+$ lying between S and S' of $N\bar{N}W_\mu$ (representing the gravitational energy flux) and of $|\phi_{AB}\eta^A\eta^B|^2W_\mu$ (representing the electromagnetic energy flux). Thus, if S' is later than S, we have $P_0 \geqq P'_0$ and $P_\mu - P'_\mu$ is a time-like vector. If neutrinos are present, $(i\nu D\nu - i\bar{\nu}D\nu)W_\mu$ gives the neutrino energy-momentum contribution, where $\nu = \phi_A\eta^A$ and D is derivative in the null direction at $\mathscr{I}^+$. Note that $DN = \phi_{ABCD}\eta^A\eta^B\eta^C\eta^D$, so that in each case the energy-momentum flux can be expressed in terms of the radiation field. However, in the spin 2 case it is the time integral of the radiation field that enters squared. In the spin $\frac{1}{2}$ case a time derivative is involved.

Lecture III

It is of interest to study the nature of infinity in the various cosmological models. This can be done using the conformal technique described in lectures I and II. The nature of the hypersurface $\mathscr{I}$ can be related to the existence of visual horizons and to the question of advanced and retarded fields in cosmological backgrounds. The space-times encountered will not generally be asymptotically simple, however.

For example, as D. Norman has pointed out to me, in the case of the Einstein–de Sitter model we necessarily have $\nabla_\mu\Omega = 0$ at $\mathscr{I}^+$. We can also represent $\mathscr{I}^-$ for this model if we allow $\Omega = \infty$ at $\mathscr{I}^-$. This is an expanding model which starts from an initial point singularity. The initial singularity can still be represented as a non-singular boundary hypersurface of $\mathscr{M}$. This is a kind of opposite procedure to that

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

considered earlier. Instead of the infinite compression ($\Omega = 0$) needed to make infinity finite, this time an infinite expansion ($\Omega = \infty$) turns a singular point into a non-singular hypersurface. This initial hypersurface turns out to be space-like. On the other hand, $\mathscr{I}^+$ turns out to be null, as in the Minkowski case, with a future vertex I^+ and a past vertex I° that lies on $\mathscr{I}^-$. The world lines of galaxies ("fundamental observers") are lines originating at points of $\mathscr{I}^-$ and terminating at I^+ (see Fig. 6).

In the steady-state model there is no initial explosion. However, in this case, $\mathscr{I}^-$ still does not represent a region properly at infinity. The model is not *complete* in the past in the sense that there are time-like goedesics which cannot be extended to infinite length into the past. Only *half* of a complete de-Sitter hyperboloid (not to

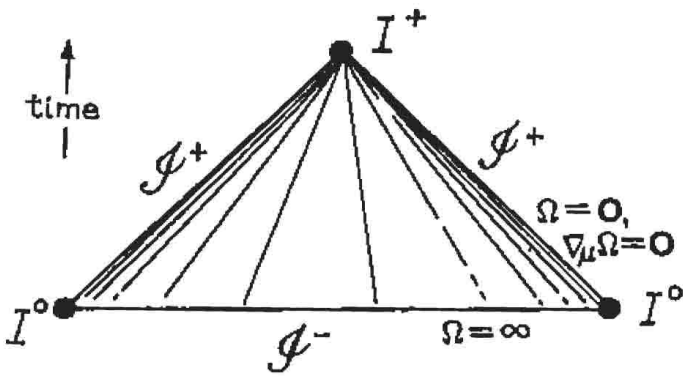


Fig. 6. $\mathscr{M}$ for the Einstein–de Sitter model ($\lambda = 0$, $k = 0$), with the world lines of the galaxies drawn in. I° is to be a single point. The initial singularity is represented by $\mathscr{I}^-$. The relativistic cosmology with $\lambda = 0$, $k < 0$ is similar except that I° is an S^2. The corresponding cases with $\lambda \neq 0$ look similar, except that $\mathscr{I}^+$ is space-like and I^+ is absent.

be confused with the Einstein–de Sitter model) is represented by the steady-state universe. The boundary $\mathscr{I}^-$ is an ordinary null hypersurface with respect to the *physical* metric. Thus we have Ω finite (non-zero) at $\mathscr{I}^-$. On the other hand, the model is asymptotically simple with respect to *future* directions: $\mathscr{I}^+$ is space-like with $\Omega = 0$, $\nabla_\mu \Omega \neq 0$. Also $\mathscr{I}^-$ has a past vertex I^- at which $\Omega = 0$ and a future vertex I° lying on $\mathscr{I}^+$. The conformal manifold $\mathscr{M}$ with boundary is, in fact, identical with that for the Einstein–de Sitter model, except that the time direction is reversed (see Fig. 7). Again, this model represents an expanding universe. The world lines of galaxies are represented (because of the continual creation) as lines originating in the interior of $\mathscr{M}$ and terminating on $\mathscr{I}^+$. There is an infinite density of these lines at $\mathscr{I}^+$ (and at $\mathscr{I}^-$) but not in the interior of $\mathscr{M}$. Each such line if extended backwards would pass through I^- (assuming no proper motions).

As with the discussion of the Minkowski case given in lecture I, each of these models can be represented conformally as a portion of the Einstein static universe (Fig. 8). In this case, just half the region representing Minkowski space-time is covered. To represent the complete asymptotically simple de Sitter space-time (contracting and expanding model), two of these portions must be joined along their null boundary—one being inverted. The result is simply a finite cylindrical section

of the infinite cylinder representing the Einstein static universe (Fig. 9). The boundary surfaces $\mathscr{I}^-$ and $\mathscr{I}^+$ are disjoint and non-singular, each being a three-dimensional sphere (S^3). In fact, this particular representation applies also to all the other relativistic (vanishing pressure) Robertson–Walker models with *positive* spatial

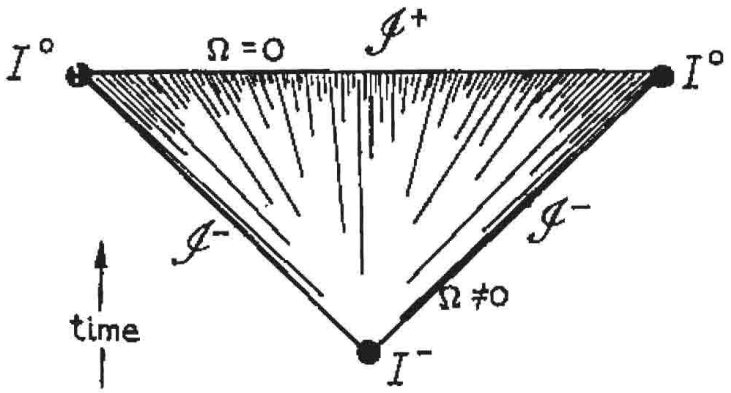


Fig. 7. $\mathscr{M}$ for the steady-state model showing the world lines of the galaxies. Again I° is a single point.

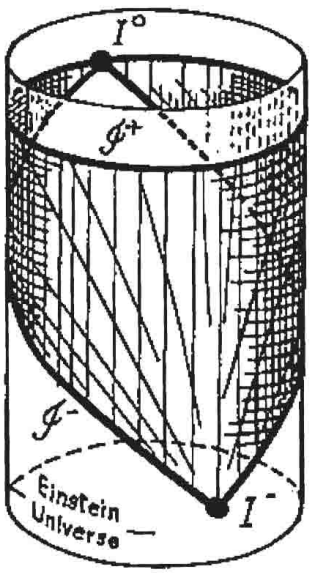


Fig. 8. The representation of the steady-state model conformally as part of an Einstein universe. This brings out that I° is a single point. The representation of the Einstein–de Sitter model is the same but upside-down. In the case of the relativistic models with $k < 0$, I° is not a point because $\mathscr{M}$ does not wrap all the way around the cylinder.

curvature k (e.g., the Lemaitre model), except that the cylinder becomes semi-infinite for the two models (e.g., Eddington–Lemaitre) which are asymptotically Einstein universes. The Einstein universe itself is a kind of *maximal* universal conformally flat space-time. That is to say, every conformally flat model is conformal to

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

a portion of the Einstein universe. The infinity of the Einstein universe cannot be represented as a non-singular hypersurface. Any conformal factor which makes the "ends" of the cylinder finite must also pinch the cylinder to a point at each end. A compression in the time direction must be accompanied by an equal compression in the space directions. Thus, there is just an initial singular point I^- and a final singular point I^+ for this case.

The Einstein–de Sitter model is the simplest expanding relativistic cosmology, since it has $k = 0$, $\lambda = 0$. The corresponding models with $k = 0$ and $\lambda < 0$ or

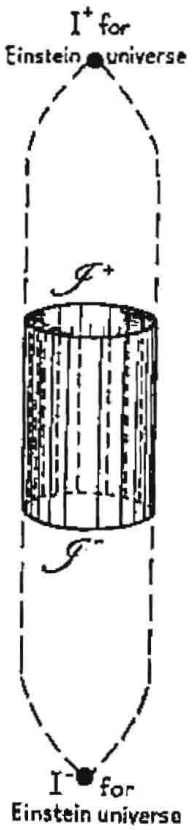


Fig. 9. The complete de Sitter space-time represented on the Einstein universe (shown dotted). This model can be obtained by joining two copies of the model shown in Fig. 8, one of which is inverted. Except for the cases which are asymptotically Einstein universes, every relativistic cosmology with $k > 0$ has a similar representation.

$\lambda > 0$ have very similar finite conformal representations except that now *both* $\mathscr{I}^+$ and $\mathscr{I}^-$ are space-like. They are joined to each other by the single point I^0. The cases with $k < 0$ correspond closely to those with $k = 0$. Their finite conformal representations differ only in that I^0 is now no longer a point but a space-like spherical surface (S^2). The Milne model's representation differs from that of Minkowski space-time in precisely the same way. When represented on the Einstein universe, unlike the cases with $k = 0$, these models do not wrap all the way around the cylinder.

In most physically interesting cases each of $\mathscr{I}^-$ and $\mathscr{I}^+$ is either space-like or null. (An example for which $\mathscr{I}$ is timelike is afforded by the de Sitter model with negative λ.) The space-like nature of $\mathscr{I}^-$ or $\mathscr{I}^+$ in many of these models is closely related to the existence of one or other of the two types of visual horizon discussed in a paper by Rindler. Suppose first that $\mathscr{I}^-$ is space-like. Now, the world lines of

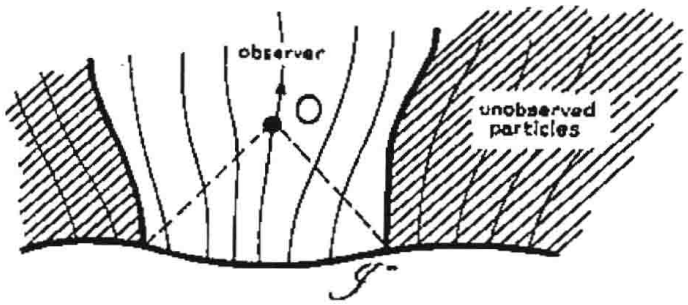


Fig. 10. When $\mathscr{I}^-$ is space-like, a particle horizon occurs. The observer can see only some of the particles in the universe. As his time progresses, more and more particles come into view.

particles or "fundamental observers" will not all meet $\mathscr{I}^-$ at the same point. (For if they did, the space-time outside the null cone of this point should not properly be considered part of the universe. $\mathscr{I}^-$ ought then to be redefined as either this null cone or possibly a time-like hypersurface within it.) Consider a particular observer and an event 0 on his world line sufficiently close to $\mathscr{I}^-$. The past null cone of 0

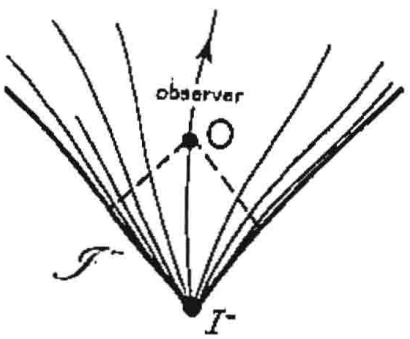


Fig. 11. No particle horizon occurs if $\mathscr{I}^-$ is null since all particles can always be seen by the observer (assuming no continual creation).

will intercept some but not all of the particles in the universe before it meets $\mathscr{I}^-$ (see Fig. 10). As 0 moves into the future along the observer's world line, more and more particles come into view. This division of the particles into those seen and those not seen is termed a *particle horizon*. On the other hand, if $\mathscr{I}^-$ is null, we expect that the world lines of the fundamental observers all pass through a past vertex I^- of $\mathscr{I}^-$. The past null cone of any event 0 near I^- clearly intercepts all these world lines and there is no particle horizon (Fig. 11).

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

Suppose, now, that $\mathscr{I}^+$ is space-like. The world line of any observer terminates at a point 0^* on $\mathscr{I}^+$. The past null cone of 0^* divides the universe into events which the observer can see at some stage of his history and those events that he can never see no matter how long he waits (Fig. 12). Suppose a clock crosses the past null cone of 0^* at some event at which the clock registers 12:00. Then, the observer sees the

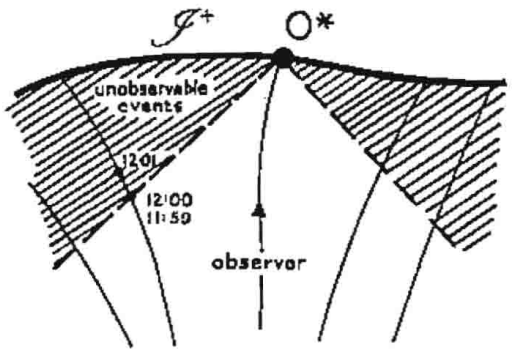


Fig. 12. An event horizon occurs when $\mathscr{I}^+$ is space-like. The events outside the past null cone of 0^* are never seen by the observer.

clock register 11:30, 11:55, 11:59, . . . , but he will never see it register 12:01. To an observer sitting on the clock, however, nothing special occurs at 12:00. This division of events into those ultimately seen and those never seen is termed an *event horizon*. If $\mathscr{I}^+$ is null, we expect the world lines of the fundamental observers all to pass through a future vertex I^+ of $\mathscr{I}^+$. In this case $0^* = I^+$ and its past null cone is

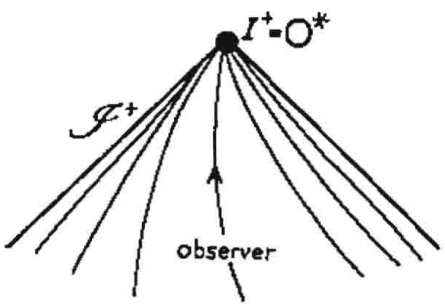


Fig. 13. No event horizon arises if $\mathscr{I}^+$ is null since the past null cone of 0^* coincides with $\mathscr{I}^+$.

the future boundary to the universe. Thus, no event horizon occurs (Fig. 13). We, therefore, have the result that:

(i) *a particle horizon corresponds to $\mathscr{I}^-$ space-like,*
(ii) *an event horizon corresponds to $\mathscr{I}^+$ space-like.*

Let us now examine the connection between these possible situations and the question of advanced and retarded fields in cosmological backgrounds. It is, of course, well known that the solutions of Maxwell's equations, consistent with a

specified charge distribution and specified motions, are non-unique up to the addition of an arbitrary solution of the source-free equations. It is usual to divide the resulting class of solutions into the (unique) retarded solution in which no waves come in from infinity in the past, the (unique) advanced solution in which no waves escape to infinity in the future, and mixed cases involving both incoming and outgoing radiation. Thus, in order to employ these ideas consistently in cosmological backgrounds, an adequate definition of the incoming and outgoing radiation fields is required. For this purpose, the ideas introduced in lecture II can be used. However, to be a little more explicit, and to motivate the discussion for the cases when $\mathscr{M}$ is not asymptotically simple, let us consider how solutions of Maxwell's equations can be built up from given initial data.

Fortunately, any isotropic space-time is conformally flat. This simplifies things immensely for the cosmological models under consideration, since we can suppose

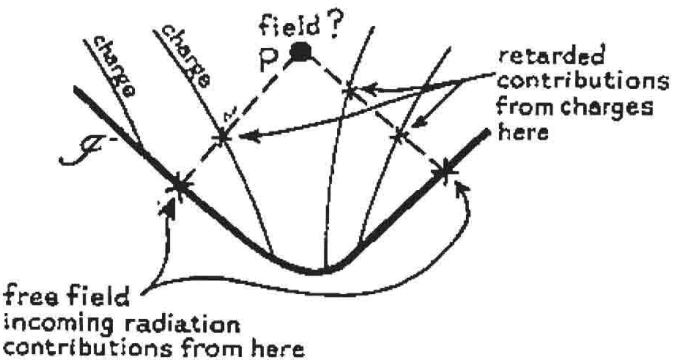


Fig. 14. If $\mathscr{M}$ is Minkowskian, we would hope to be able to express the Maxwell field at p in terms of retarded contributions from the charges intercepted by the past null cone of p, plus a free-field contribution arising from the field on $\mathscr{I}^-$.

that the metric of $\mathscr{M}$ is (near $\mathscr{I}^-$ or $\mathscr{I}^+$ at least) just that of flat Minkowski space-time. Let P be a point of $\mathscr{M}$ not too far from $\mathscr{I}^-$, say. Then we may expect to be able to express the (conformally invariant) field at P as a sum of contributions from the charges arising from the various intersection points of the charges' world lines with the past null cone of P, together with an integral over the intersection of $\mathscr{I}^-$ with this null cone (Fig. 14). The field at P would be a sum of the "retarded" fields due to the charges and a free field contribution arising from $\mathscr{I}^-$—the incoming radiation field. A problem here is how to express the retarded field of the charges. One way to do this might be to find a particular solution of Maxwell's equations for each charge and to subtract out any free field contribution arising from $\mathscr{I}^-$. The matter clearly rests on the behaviour of the *free field* in relation to its values on the hypersurface $\mathscr{I}^-$.

For increased generality, let us consider a free field of arbitrary spin s in Minkowski space-time. As in lecture I, we can represent this by a symmetric spinor $\phi_{A\ldots K}$ with $2s$ indices subject to $\nabla^{AA'}\phi_{AB\ldots K} = 0$ (or $\nabla_\mu \nabla^\mu \phi = 0$ if $s = 0$). The

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

field at point P can be expressed as an integral over quantities defined on the intersection Σ of the null cone of P and an initial hypersurface $\mathscr{S}$:

$$\phi_{AB\ldots K}(P) = \frac{1}{2\pi}\int \frac{1}{p}\zeta_A\zeta_B\ldots\zeta_K\{D\phi - (2s+1)\rho\phi\}\,d\Sigma.$$

If Q denotes the general point of Σ, the various quantities appearing above are defined as follows (see Fig. 15). The spinor equivalent of the 4-vector $\overrightarrow{QP}$ is $p\zeta_A\bar{\zeta}_B$, so that the spinor ζ_A corresponds to the null direction of $\overrightarrow{QP}$ and p is a measure of the extent of $\overrightarrow{QP}$. The element of surface area at Q is $d\Sigma$. Let the spinor ξ^A be defined at Q such that the corresponding null direction is coplanar with that of ζ_A and the normal n^μ of $\mathscr{S}$ at Q. That is to say, we have $\xi^A\xi^{\dot{B}}$ linearly dependent on $n^{A\dot{B}}$

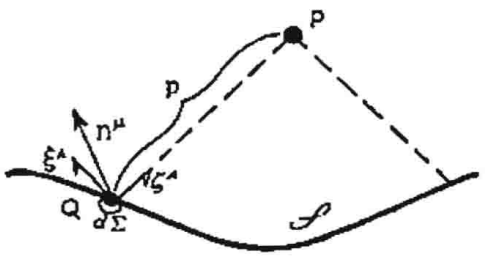


Fig. 15. In Minkowski space-time, the free field at P can be simply expressed in terms of quantities defined on an initial hypersurface $\mathscr{S}$. The relevant field component (and its derivative) that is required on $\mathscr{S}$ is the one contracted completely with ξ^A. The normal $n\mu$ to $\mathscr{S}$ at Q is coplanar with the null direction of ξ^A and ζ^A, where ζ^A points along QP. Thus, only when $\mathscr{S}$ is null can ξ^A be independent of the position of P.

$(= n^\mu\sigma_\mu{}^{A\dot{B}})$ and $\zeta^A\zeta^{\dot{B}}$. The normalization of ξ^A is to be fixed by $\zeta_A\xi^A = 1$. The complex quantity ϕ is defined by

$$\phi = \phi_{AB\ldots K}(Q)\xi^A\xi^B\ldots\xi^K.$$

The expression $D\phi$ denotes the derivative of this component of $\phi_{A\ldots K}$ in the null direction corresponding to ξ^A, i.e., $D = \xi^A\xi^{\dot{B}}\nabla_{A\dot{B}}$. The quantity $-\rho$ measures the divergence of these null directions at Q, i.e., $\rho = \zeta_A\bar{\xi}_{\dot{B}}\xi_C\nabla^{A\dot{B}}\xi^C$.

Comparing this expression for ϕ with some results given in lecture II, we see that if $\mathscr{S} = \mathscr{I}^-$ and the space-time is asymptotically simple, then ϕ measures the incoming radiation field as defined there. If $\mathscr{S} = \mathscr{I}^+$ we get the outgoing radiation field. In these cases this is the r^{-1} part of the field, where r is an affine parameter on a null geodesic through P. If the space-time fails to be asymptotically simple at $\mathscr{I}^+$ or $\mathscr{I}^-$ because $\nabla_\mu\Omega = 0$ there, then D. Norman has pointed out that the affine parameter r need only be replaced by a luminosity parameter r^* ($\sim\Omega^{-1}$) and the peeling-off property still holds. Thus ϕ measures the r^{*-1} part of the field. If $\Omega \neq 0$ (e.g., $\Omega = \infty$) at $\mathscr{I}^-$ or $\mathscr{I}^+$, then the interpretation of ϕ as measuring the radiation field is not quite so direct in these terms. But generally, it is the fact that ϕ is the quantity appearing in the free-field integral for $\phi_{A\ldots K}$, which suggests such an interpretation.

Note added in proof. Replace ξ, ζ and η, ξ to get notational agreement with Lecture II.

One important property of the quantity ϕ should be pointed out at this stage. If the initial hypersurface $\mathscr{S}$ is null, then at any point Q of $\mathscr{S}$, ξ^A is effectively independent of the position of P, since the direction of ξ^A is the null direction in $\mathscr{S}$—which is the normal to $\mathscr{S}$. On the other hand, if $\mathscr{S}$ is space-like, then the direction of ξ^A depends on the direction of $\mathfrak{Z}^A$. That is, ξ^A depends essentially on the position of P. The radiation field ϕ is therefore an origin dependent concept when $\mathscr{S}$ is space-like, but it is an origin independent concept when $\mathscr{S}$ is null.

Suppose we require that the incoming radiation field be zero with respect to *all* origins (purely retarded field). Then if $\mathscr{S}^-$ is null this represents the one complex equation $\phi = 0$ on $\mathscr{S}^-$. If, on the other hand, $\mathscr{S}^-$ is space-like, we must have all components vanishing: $\phi_{A\ldots K} = 0$ on $\mathscr{S}^-$. This is $2s + 1$ complex equations. However, on a space-like hypersurface only four real variables are specifiable arbitrarily as initial data for $\phi_{A\ldots K}$. Thus, if $s > 1/2$ and given sources are present, we cannot in general expect to arrange $\phi_{A\ldots K} = 0$ on $\mathscr{S}^-$. On a null hypersurface, the two real

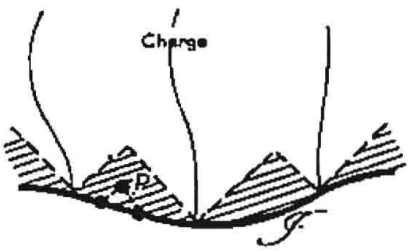


Fig. 16. If $\mathscr{S}^-$ is space-like and the charges meet it in a discrete set of points, then there will be regions near $\mathscr{S}^-$ of points P whose past null cones do not meet the world lines of the charges at all. If incoming field is absent, we would expect the field at P to vanish. However, this leads to an inconsistency.

quantities defined by ϕ can be specified arbitrarily as initial data for $\phi_{A\ldots K}$. Thus, if $\mathscr{S}^-$ is null, $\phi = 0$ can always be arranged on $\mathscr{S}^-$ whatever the source distribution. The result is that the incoming radiation field can be made to vanish, for general source distributions and with respect to all origins, if and ($s > \frac{1}{2}$) only if $\mathscr{S}^-$ is null. The corresponding result for $\mathscr{S}^+$ and the outgoing radiation field also holds.

This negative result for $\mathscr{S}^-$ space-like is perhaps brought out more forcefully if we consider a discrete set of charges and those points P sufficiently near $\mathscr{S}^-$ whose past null cones do not intersect the world line of any charge (see Fig. 16). If the incoming radiation (as defined here) is to vanish, then the Maxwell field at P must vanish. We can surround the past end of a charge by such points P. But the (conserved) charge can be measured by an integral of the Maxwell field over any sphere surrounding the world line of the charge. Therefore, the Maxwell field *cannot* vanish at all such points P and incoming radiation must be present. To sum up, we have the following result:

(iii) *if there is a particle horizon, then purely retarded fields of spin* $> \frac{1}{2}$ *do not exist for general source distributions.*

(iv) *if there is an event horizon, then purely advanced fields of spin* $> \frac{1}{2}$ *do not exist for general source distributions.*

 Conformal treatment of Infinity. Relativité, groupes et topologie. *Relativity, groups and topology.* Lectures delivered at Les Houches during the 1963 session of the Summer School of Theoretical Physics, University of Grenoble. Ed. DeWitt & DeWitt. New York

It should be emphasized that the above result depends on the definition of advanced and retarded fields given here. It is conceivable that an alternative definition might be found which leads to different results. Also, the application of the result to actual physical models is not at all clear cut. For example, if in an exploding model, all the matter is considered to be in the form of neutral particles throughout the initial stages of its existence, then purely retarded fields could be possible. The consideration of the spin 2 case might be of interest in this context, since gravitationally neutral particles are impossible. (A spin 2 field ϕ_{ABCD} unrelated to the curvature of $\tilde{\mathscr{M}}$ represents a weak field gravitational perturbation of $\tilde{\mathscr{M}}$.)

The question of asymptotic symmetry groups for $\tilde{\mathscr{M}}$ is another which relates to the geometry of the hypersurface $\mathscr{I}$. Any motion which takes $\tilde{\mathscr{M}}$ into itself will clearly correspond to a conformal motion of $\mathscr{M}$ into itself. Hence, there is induced a conformal motion of the boundary $\mathscr{I}$ into itself. For a general space-time $\tilde{\mathscr{M}}$ with a detailed structure there will be no exact motions taking $\tilde{\mathscr{M}}$ into itself. However, the conformal motions of $\mathscr{I}$ into itself may persist. This gives an asymptotic symmetry group. When $\mathscr{I}^+$ is space-like it is an ordinary conformal Riemannian 3-manifold, so the question as to what are its conformal motions is a straightforward one. However, for asymptotically flat space-times, for example, $\mathscr{I}$ is null. The intrinsic metric of $\mathscr{I}$ has vanishing determinant. Consequently, the concept of the conformal geometry of $\mathscr{I}$ admits more than one possible interpretation. The intrinsic conformal metric of $\mathscr{I}$ alone seems to give it too little structure to give rise to a physically interesting symmetry group. However, the conformal geometry of $\mathscr{I}$ can be strengthened if certain conformally invariant concepts are taken over from the imbedding of $\mathscr{I}$ in $\mathscr{M}$. This will be discussed here for the case of an asymptotically flat space-time $\tilde{\mathscr{M}}$ only.

One way of defining a conformal transformation is to say that it is a transformation which preserves *angles*. Let us apply this idea to the conformal geometry of the null hypersurface $\mathscr{I}^+$ for an asymptotically flat space-time. There are two types of angles to be distinguished on $\mathscr{I}^+$, namely, ordinary finite angles and *null angles*. A finite angle arises between two distinct directions in $\mathscr{I}^+$, at a point on $\mathscr{I}^+$, which are not coplanar with the null direction ι in $\mathscr{I}^+$. Two directions at a point of $\mathscr{I}^+$, different from ι, which *are* coplanar with ι define a null angle on $\mathscr{I}^+$. The numerical value of a null angle is necessarily zero, but nevertheless it makes sense to say that one null angle is greater than another null angle if it includes the other null angle as a part. To make this idea globally consistent it is necessary to establish an equivalence between null angles over the whole of $\mathscr{I}^+$. This can be done in the following way. Choose a particular metric for $\mathscr{M}$ such that the divergence of $\mathscr{I}^+$ vanishes. This is always possible when $\tilde{\mathscr{M}}$ is asymptotically flat, for then the shear of $\mathscr{I}^+$ must also vanish. The shear and divergence of $\mathscr{I}^+$ both vanishing, it is possible to define parallel transport of vectors within $\mathscr{I}^+$ by appealing to the Christoffel symbols defined for $\mathscr{M}$. This notion of parallel transport of vectors defines parallel transport of null angles. Two properties of this parallel transport can now be proved. Firstly, it establishes an *integrable equivalence* between the null angles and secondly, this equivalence is independent of the particular metric chosen, as above, for $\mathscr{M}$. Thus, a *strong conformal geometry* for $\mathscr{I}^+$ is established. The self-transformations of $\mathscr{I}^+$ which preserve this strong conformal geometry (and do not reverse time sense) define the *Bondi–Metzner–Sachs asymptotic symmetry group* for asymptotically flat

space-times. (A definition of what is meant by "asymptotically flat" has not been precisely given here. If $\mathscr{M}$ is asymptotically simple and $R_{\mu\nu} = 0$ in the neighbourhood of $\mathscr{I}$, this will suffice. Also, the inclusion of reasonable electromagnetic fields near $\mathscr{I}$ does not essentially change this situation.)

When $\mathscr{M}$ is Minkowski space-time it is possible to single out the inhomogeneous Lorentz subgroup of the Bondi–Metzner–Sachs group. This can be done by choosing only those self-transformations of $\mathscr{I}^+$ (preserving strong conformal geometry) which are regular at I^+ or, equivalently, at I°. This rules out the "supertranslations" characteristic of the full B.M.S. group, in which the various generators of $\mathscr{I}^+$ can be "translated" independently. If $\mathscr{M}$ is only asymptotically flat, then I^+ may be singular. We would expect this to be the case whenever matter is present at future infinity. That is, I^+ would be non-singular only if all the matter ever present in the system is ultimately radiated away in the form of zero rest-mass fields. The situation is even worse as regards I°. For I° is singular for Schwarzschild's solution, so it would appear that it is singular for *any* asymptotically flat system possessing mass. Thus, it seems that the inhomogeneous Lorentz group does not, in general, emerge as an asymptotic symmetry group for asymptotically flat space-times.

References

Bateman, H. 1910, *Proc. Lond. Math. Soc.* **8**, 223.
Bergmann, P. G. 1961, *Phys. Rev.* **124**, 274.
Bondi, H. 1952, *Cosmology*, Cambridge.
——, van der Burg, M. G. J. and Metzner, A. W. K. 1962, *Proc. Roy. Soc.* A **269**, 21.
Buchdahl, H. A. 1959, *Nuovo Cimento* **11**, 496.
Cunningham, E. 1910, *Proc. Lond. Math. Soc.* **8**, 77.
Fierz, M. 1940, *Helv. Phys. Acta.* **13**, 45.
Hogarth, J. E. 1962, *Proc. Roy. Soc.* **A267**, 365.
McLennan, J. A. (Jr.) 1956, *Nuovo Cimento* **10**, 1360.
Newman, E. T. and Penrose, R. 1962, *J. Math. Phys.* **3**, 566.
—— and Unti, T. W. J. 1962, *J. Math. Phys.* **3**, 891.
Penrose, R. 1960, *Ann. of Phys.* **10**, 171.
—— "Null Hypersurface Initial Data for Classical Fields of Arbitrary Spin and for General Relativity," preprint 1961 (to be published in extended form).
—— 1962, *Report to the Warsaw Conference* (to appear).
—— 1963, *Phys. Rev. Lett.* **10**, 66.
—— 1963, *Report to the Cornell "Nature of time" Meeting* (to appear).
Rindler, W. 1956, *Month. Nat. Roy. Astr. Soc.* **116**, 662.
Robinson, I. and Trautman, A. 1962, *Proc. Roy. Soc.* A **265**, 463.
Sachs, R. K. 1961, *Proc. Roy. Soc.* A **264**, 309.
—— 1962, *Proc. Roy. Soc.* A **270**, 103.
—— 1962, *Phys. Rev.* **128**, 2851.
Schrodinger, E. 1956, *Expanding Universes*, Cambridge.
Sciama, D. W. 1963, *Proc. Roy. Soc.* A **273**, 484.
Trautman, A. 1958, *Bull. Acad. Polon. Sci.* **6**, 403, 407.

19
ENERGY CONSERVATION AS THE BASIS OF RELATIVISTIC MECHANICS

Am. J. Phys. **33(1)** (1965), 55–59. With W. Rindler

See also Chapter 24.

Energy Conservation as the Basis of Relativistic Mechanics

R. PENROSE*
University of Texas, Austin, Texas
AND
W. RINDLER
Southwest Center for Advanced Studies, Dallas, Texas
(Received 22 June 1964)

The relativistic theory of collisions of macroscopic particles is developed from the two axioms of energy conservation and relativity, by use of standard relativistic kinematics (without, of course, assuming the mathematical expression for relativistic energy). We derive, in turn, the equivalence of rest-mass and rest-energy, the usual mathematical expression for the total energy in terms of the rest-energy, the conservation of 3-momentum, and the equivalence of total energy with inertial mass. We incidentally find all possible conservation laws of quantities described by the product of a scalar with a function of the speed.

IT is quite usual to base the development of special relativistic mechanics on *two* conservation principles: those of linear momentum and of inertial mass.[1] Nevertheless, it is known that the conservation of inertial mass can be *deduced* from the conservation of momentum in conjunction with the relativity principle[2] (momentum being defined as *some* scalar function of velocity times the 3-velocity vector, and inertial mass as that scalar function.) An even more fundamental starting point is the energy principle. Both Weyl[3] and Costa de Beauregard[4] mention that P. Langevin, in "elegant" lectures delivered around 1922 in Zurich and at the Collège de France, developed relativistic mechanics assuming only energy conservation and the relativity principle (without, of course, assuming the mathematical expression of the relativistic energy—for once that is known, the rest follows trivially.) Costa de Beauregard says that these lectures were, unfortunately, not published and also implies that they were not simple. Some time after completing our attempt to reconstruct Langevin's argument (as presented below) we unearthed what appears to be an outline of Langevin's original scheme. This is contained in a 1932 lecture by F. Perrin entitled "La Dynamique Relativiste et l'Inertie de l'Énergie,"[5] which Langevin himself edited for publication. Though his and our arguments are within the same general area of ideas, ours appears to be mathematically considerably simpler. Also, whereas some of the detailed arguments for key assumptions are necessarily omitted from Perrin's brief report, we had tried to justify ours as completely as possible. We were unable to discover any reference to the possibility of such a development of relativistic mechanics in any text published in English. In view of this situation, we thought it worth while to present our version of the argument.

We concern ourselves here only with the theory of collisions of macroscopic particles. Between collisions, the particles are assumed to experience no forces and thus, in accordance with Newton's first law, move uniformly relative to any inertial frame. The forces acting during

* Present address: Birkbeck College, London, England.

[1] See, for example, G. N. Lewis and R. C. Tolman, Phil. Mag. 18, 510 (1909); R. C. Tolman, Phil. Mag. 23, 375 (1912); A. S. Eddington, *The Mathematical Theory of Relativity* (Cambridge University Press, Cambridge, England, 1923), p. 31; R. C. Tolman, *Relativity, Thermodynamics and Cosmology* (Oxford University Press, Oxford, England, 1934), p. 42; W. H. McCrea, *Relativity Physics* (Methuen and Company, Ltd., London, 1935), p. 19; J. L. Synge, *Relativity: The Special Theory* (North-Holland Publishing Company, Amsterdam, 1956), p. 174; W. Rindler, *Special Relativity* (Oliver and Boyd, Edinburgh, 1960), p. 76.

[2] This is done, for example, by H. Weyl in Sec. 27 of the 5th (German) edition of *Raum, Zeit, Materie* (Julius Springer Verlag, Berlin, 1923), though not in the only available English version, which is a translation of the 4th German edition; and by C. Møller in his *Theory of Relativity* (Oxford University Press, Oxford, England, 1952), Secs. 26–28.

[3] See H. Weyl, Ref. 2, p. 203.

[4] O. Costa de Beauregard, *La Théorie de la relativité restreinte* (Masson et Cie, Paris, 1949), p. 87.

[5] F. Perrin, *Actualités scientifiques et industrielles* (Hermann et Cie, Paris, 1932), Serie 1932, Vol. XLI.

the "collision" are regarded either as short-range or contact forces.

By analogy with classical mechanics, we assume that the total energy of a particle is made up of (i) kinetic (or external) energy, and (ii) rest- (or internal) energy. Kinetic energy is that which a particle possesses in virtue of being in motion relative to the observer's inertial frame, and which can be fully extracted by letting the particle come to rest in an encounter with an absorbing system (e.g., a spring or a second particle). Rest-energy, on the other hand, is that which can be extracted from the particle in its rest-frame; it may exist in the form of heat, elastic tension, chemical energy, etc.; it is a scalar associated with the particle and is, by definition, measured in the rest-frame.

In classical mechanics, the kinetic energy of a "particle" is proportional to the square of its velocity and to the "amount of matter" in it. In relativity, we must also be prepared to assume that there could be associated with each particle an "amount of matter," which affects its kinetic energy, but which is itself *not* a form of energy. This could be, for example, some linear combination of the numbers of constituent electrons, protons, neutrons, etc. If we assume that this quantity has independent existence, we should also assume that it remains constant for each particle between collisions. We further allow that rest-energy, too, may contribute to the kinetic energy. If the presence of *any* form of rest-energy in a particle affects its kinetic energy, then *all* forms of rest-energy must affect kinetic energy equally, since all forms of energy are presumed to be convertible into each other. We suppose, then, that the "amount of matter" associated with each particle is measured by a scalar m; that rest-energy is measured by another such scalar μ; and that m and μ remain constant between collisions. Since the velocity of each particle also remains constant between collisions and since, because of the isotropy of space, only the magnitude u of the velocity can enter the expression for the kinetic energy, the most general functional form which the kinetic energy can take is

$$K = K(m,\mu,u). \tag{1}$$

We could even allow for the possibility that different kinds of matter (e.g., protons, neutrons, etc.) affect the kinetic energy differently, and accordingly set $K = K(m_1,m_2, \ldots,\mu,u)$, where m_1 is the "proton matter," m_2 the "neutron matter," etc. This would not, however, substantially alter our arguments nor affect our conclusions.

We now require that two particles traveling side by side have the same kinetic energy as a single particle whose m and μ are the sums of the corresponding quantities of the separate particles. In other words, we require *additivity* of kinetic energy:

$$K(m_1,\mu_1,u)+K(m_2,\mu_2,u) = K(m_1+m_2,\ \mu_1+\mu_2,\ u). \tag{2}$$

If $K(m,\mu,u)$ is a continuous function of its arguments, as we may reasonably require; and if m and μ can vary independently, as is the case for macroscopic particles (e.g., they can be heated); then condition (2) implies that K must be of the following form[6]

$$K(m,\mu,u) = mf(u)+\mu g(u), \tag{3}$$

where $f(u)$ and $g(u)$ are functions of u alone, i.e., universal functions independent of the particle, which are evidently continuous because of the assumed continuity of K.

For convenience we introduce as a new variable the "Lorentz factor"

$$\gamma = \gamma(u) = (1-u^2/c^2)^{-\frac{1}{2}}, \tag{4}$$

and we write $f(u) \equiv F(\gamma)$, $g(u) \equiv G(\gamma)$. These functions are clearly continuous in γ. Equation (3) now becomes

$$K(m,\mu,u) = mF(\gamma)+\mu G(\gamma). \tag{5}$$

By definition, kinetic energy is fully spent when a particle is brought to rest. Hence, we must

[6] The familiar proof runs as follows: let $\phi(\mathbf{x})$ be a scalar valued continuous function of the vector $\mathbf{x}=(x_1,x_2)$. Suppose $\phi(\mathbf{x}+\mathbf{y})=\phi(\mathbf{x})+\phi(\mathbf{y})$; then clearly $\phi(n\mathbf{x})=n\phi(\mathbf{x})$ (n any integer); put $n\mathbf{x}=\mathbf{y}$: $\phi((1/n)\mathbf{y})=(1/n)\phi(\mathbf{y})$; thus $\phi((n/m)\mathbf{x})=(n/m)\phi(\mathbf{x})$ (n, m any integers) and, by continuity, $\phi(\lambda\mathbf{x})=\lambda\phi(\mathbf{x})$ (λ any real nonnegative number); write $\phi((1,0))=a_1$, $\phi((0,1))=a_2$; then, if x_1, $x_2 \geq 0$, $\phi(\mathbf{x})=a_1x_1+a_2x_2$. This proof is easily generalized to the case where x_1 (or x_2, or both) can take positive and negative values, or negative values only. In the former cases we have $\phi((-x_1,x_2))=\phi((0,2x_2))-\phi((x_1,x_2))=-a_1x_1+a_2x_2$; in the latter case we would define $\phi((-1,0))=a_1$. In the text, K depends not only on the "vector" $\mathbf{x}=(m,\mu)$ but also on the "parameter" u. Since the above proof applies for each fixed u, we get Eq. (3).

have $mF(1)+\mu G(1)=0$; but since m and μ are independent, this implies

$$F(1)=0, \quad G(1)=0. \tag{6}$$

We next determine the functional forms of F and G from the following assumptions: (i) energy is conserved in all collisions, and (ii) this is true in all inertial frames (relativity principle). To this end, consider the following rather contrived collision between two identical particles with associated scalars m and μ. In an inertial frame S the particles approach each other with equal and opposite velocities u along the positive and negative x axis, collide at the origin, and then separate with the same equal and opposite velocities u along the y axis, and with the same m and μ as before. Very few assumptions are needed to justify the feasibility of such a collision. If we grant the possibility of head-on collisions (directions of motion reversed) and of grazing collisions (directions of motion practically unchanged) then, by continuity and symmetry there will be collisions with the directions changed by a right angle. By symmetry, furthermore, the particles must separate with equal speeds, say v. Finally, if we grant the possibility of "elastic" collisions (m and μ of each particle unchanged), the conservation of energy implies that the original and final speeds are equal; for even if two or more distinct speeds correspond to the same kinetic energy, continuity with the grazing collision ensures $u=v$.

Now consider the same collision as viewed from a second frame S' moving with velocity u' along the x axis of S. We recall[7] that the 4-velocity U^i associated with a particle having 3-velocity $\mathbf{u}$ is given by $U^i=\gamma(u)(\mathbf{u},c)$; hence the 4-velocity of an inertial observer relative to his own rest-frame is $U_0{}^i=(0,c)$. Thus

$$\gamma=(U_iU_0{}^i)/c^2, \tag{7}$$

i.e., the Lorentz factor of a particle relative to any inertial observer is given simply by the scalar product of the two corresponding 4-velocities divided by c^2. And this product, being invariant, can be evaluated in any inertial frame. In S the 4-velocities of the two particles, before and after collision, respectively, are

$$\gamma(\pm u, 0, 0, c); \quad \gamma(0, \pm u, 0, c) \tag{8}$$

and the 4-velocity of S' is

$$\gamma'(u',0,0,c), \quad \gamma'=(1-u'^2/c^2)^{-\frac{1}{2}}. \tag{9}$$

Hence, the various Lorentz factors of the particles relative to S' are given by the scalar products (divided by c^2) of (9) with the 4-velocities in (8):

$$\gamma_1=\gamma\gamma'(1+uu'/c^2), \quad \gamma_2=\gamma\gamma'(1-uu'/c^2),$$
$$\gamma_3=\gamma_4=\gamma\gamma'=\tfrac{1}{2}(\gamma_1+\gamma_2). \tag{10}$$

Now since total energy is conserved relative to S', and since rest-energy is here conserved separately, kinetic energy must be conserved also. Thus, from (3) and (10),

$$mF(\gamma_1)+\mu G(\gamma_1)+mF(\gamma_2)+\mu G(\gamma_2)$$
$$=2mF(\tfrac{1}{2}(\gamma_1+\gamma_2))+2\mu G(\tfrac{1}{2}(\gamma_1+\gamma_2)). \tag{11}$$

This equation must balance in its F and G terms *separately*, since μ can be varied independently of m. We are, therefore, led to the functional equations

$$\left.\begin{aligned} F(\gamma_1)+F(\gamma_2)&=2F(\tfrac{1}{2}(\gamma_1+\gamma_2))\\ G(\gamma_1)+G(\gamma_2)&=2G(\tfrac{1}{2}(\gamma_1+\gamma_2))\end{aligned}\right\}, \tag{12}$$

and we fix our attention on the first of these. The linearity of its solution is graphically evident as soon as we translate the equation into words: "the average of F at γ_1 and γ_2 equals F at the average of γ_1 and γ_2." We can also find the solution formally by defining an auxiliary function $\bar{F}$ by the equation

$$F(\gamma)=\bar{F}(\gamma-1) \tag{13}$$

and substituting this into (12) (i). We find at once that $\bar{F}$ satisfies the same equation:

$$\bar{F}(x_1)+\bar{F}(x_2)=2\bar{F}(\tfrac{1}{2}(x_1+x_2)), \tag{14}$$

where $x_1=\gamma_1-1$, $x_2=\gamma_2-1$. Putting $x_2=0$ and observing that $\bar{F}(0)=0$ because of (6), we find

$$\bar{F}(x)=2\bar{F}(x/2),$$

whence (14) can be written

$$\bar{F}(x_1)+\bar{F}(x_2)=\bar{F}(x_1+x_2). \tag{15}$$

The physics of the situation requires that (12) must hold for all $\gamma_1, \gamma_2>1$ (arbitrary collision

[7] See, for example, W. Rindler, in Ref. 1, Eq. (4.14) and exercise (9), p. 39.

velocity, arbitrary velocity of S'). Hence (15) must hold for all x_1, $x_2>0$, and thus, by continuity, for all x_1, $x_2\geqslant 0$. An argument similar to the one which yielded (3) now shows that, for some constant a,

$$\bar{F}(x)=ax$$

whence, from (13),

$$F(\gamma)=a(\gamma-1). \tag{16}$$

Similarly, from (12) (ii) and for some other constant b,

$$G(\gamma)=b(\gamma-1).$$

Substituting these solutions in (5), we get

$$K=(am+b\mu)(\gamma-1)=c^2M(\gamma-1), \tag{17}$$

where we have introduced the quantity

$$M=c^{-2}(am+b\mu). \tag{18}$$

Now, in the limit of small velocities, relativistic mechanics should reduce to the well-proven classical mechanics in which $K=\text{mass}\times\frac{1}{2}u^2$. Since, by reference to (4), Eq. (17) can be written in the form

$$K=\tfrac{1}{2}Mu^2\{1+O(u^2/c^2)\},$$

we see that M is precisely the quantity recognized as the mass of the particle in a frame in which its velocity is infinitesimal. For this reason, M is called the rest-mass of the particle. Equation (18) shows that m and μ contribute in the same manner to the rest-mass, unless $a=0$ or $b=0$. In fact, as we now prove, $b=1$.

For this purpose, we consider a completely inelastic collision: the identical particles approach each other as before, but on collision unite. From symmetry, it is evident that the resulting compound particle is at rest relative to S. The first two of Eqs. (10) still apply, while the Lorentz factor of the compound particle relative to S' is simply γ'. We assume that it is possible to arrange *at least one* such collision in which the total "amount of matter" is conserved, so that it is 2m for the compound particle. Let $\bar{\mu}$ be its rest-energy. Conservation of energy in S then requires [cf. (17)]

$$\bar{\mu}=2\mu+2c^2M(\gamma-1), \tag{19}$$

while conservation of energy in S' requires

$$2c^2M(\gamma\gamma'-1)+2\mu=(2am+b\bar{\mu})(\gamma'-1)+\bar{\mu}. \tag{20}$$

If we substitute for $\bar{\mu}$ from (19) into (20), we easily find that the resulting equation can be reduced to the form

$$(\gamma-1)(\gamma'-1)=b(\gamma-1)(\gamma'-1).$$

Since $\gamma\neq 1$ and γ' need not be 1, it follows that $b=1$, as asserted. Consequently,

$$c^2M=\mu+am. \tag{21}$$

It is at this stage in *any* development of relativistic mechanics that a logical gap has to be bridged with a hypothesis. As (21) shows, any *increase* in rest-energy causes an increase in rest-mass. This "matter-like" behavior of energy strongly suggests that *all* of the rest-mass is due to internal energy, and that what we have called "amount of matter" has no separate existence but is really a form of rest-energy and should be included with μ. We are thus led, as Einstein was, to the equation

$$c^2M=\mu. \tag{22}$$

The fact that by now all elementary particles have been observed in reactions (e.g., with antiparticles) in which they get annihilated and either directly or indirectly converted into pure radiation (photons or neutrinos), today reassures us that *all* matter is totally convertible into energy and that (22) is indeed the correct form of (21).

From this point on, the development is straightforward. The total energy E of a particle is now given by

$$E=K+\mu=\mu(\gamma-1)+\mu=\mu\gamma=c^2M\gamma, \tag{23}$$

and, thus, conservation of energy through any collision requires

$$\sum M\gamma\ [\text{before impact}]=\sum{}'M\gamma\ [\text{after impact}], \tag{24}$$

the summation being extended over all the participating particles. Consider, then, the 4-vector

$$\sum MU^i[\text{before impact}]-\sum{}'MU^i[\text{after impact}].$$

Its fourth component, containing terms of type $M\gamma c$, vanishes by virtue of (24). It follows at once, from the usual transformation properties, that this whole 4-vector must vanish. Its spatial part contains terms of type $M\gamma\mathbf{u}$, whose conservation is now established:

$$\sum M\gamma\mathbf{u}[\text{before impact}] = \sum' M\gamma\mathbf{u}[\text{after impact}]. \quad (25)$$

This is analogous to the classical momentum conservation law, and so we naturally adopt $M\gamma\mathbf{u}$ as the definition of *relativistic 3-momentum*. (We choose $M\gamma\mathbf{u}$ rather than a multiple of it, because once again we wish to conform to the general requirement that relativistic mechanics should reduce to classical mechanics in the limit of small velocities.) Since momentum conservation defines *inertial mass*,[8] we must take $M\gamma$ as the inertial mass of a particle. Eq. (23) then tells us that total energy is a universal multiple (by a factor c^2) of inertial mass; hence the two are conserved together. The analogs of the separate classical laws of conservation of mass, energy, and momentum, are therefore all contained in the single relativistic law of the conservation of *4-momentum* MU^i.

We have now reached the point of confluence of the various standard developments of relativistic collision mechanics, and we need proceed no further. However, one corollary of the above work may yet be noted: in solving one of Eqs. (12) we have, in fact, found *all possible* relativistic conservation laws of the form

$$\sum sf(u)[\text{before impact}] = \sum' sf(u)[\text{after impact}], \quad (26)$$

where the s are scalars associated with the various particles, and $f(u)$ is a universal function of the magnitude of the velocity. For any such law must apply, in particular, to the type of collision leading to (12) (i). The general solution of that equation is the linear function $F(\gamma) = a\gamma + b$. However, if inelastic collisions are included, it is not hard to see [using (10)] that either a or b must vanish. Hence, there are only two solutions of Eq. (26), namely, $f(u) = b$ and $f(u) = a(1-u^2/c^2)^{-\frac{1}{2}}$. From the standard theory we know physical quantities which obey such conservation laws: in the first class are electric charge, baryon number, etc.; in the second class are energy and inertial mass.

We may mention, without proof, the curious fact that when all particles and observers are confined to one spatial dimension the class of mathematically possible conservation laws of form (26) is widened. It is given by $f(u) = a\cosh b\alpha$, where $\alpha = \cosh^{-1}\gamma$ ("rapidity") and a, b are two constants of which the former is real and the latter is real or purely imaginary. That no other laws are possible can be seen by an examination of the inelastic collision of equal particles. To show that *all* such laws are possible it is sufficient to examine an inelastic collision of unequal particles since elastic or semielastic collisions can be compounded of two inelastic ones of which the second proceeds in reverse.

[8] See, for example, W. Rindler, in Ref. 1, Sec. 28.

20
GRAVITATIONAL COLLAPSE AND SPACE-TIME SINGULARITIES

Phys. Rev. Letters **14(3)** (1965), *57–59*

This article gives a brief account of the first singularity theorem, where the notion of a 'trapped surface' is introduced to characterize a gravitational collapse that has reached a point of no return, this characterization being of a 'generic' nature. A sketch is provided of a proof that the presence of a trapped surface (together with various natural assumptions) implies that a singularity will arise. An awkwardness arises in the argument, as provided here, where it is required that the boundary 3-surface B^3 be approximated by a smooth spacelike 3-surface $B^{3}\star$, the (timelike) geodesic normals to $B^{3}\star$ providing the needed mapping to the initial Cauchy hypersurface. The much simpler argument given in Chapter 28 takes advantage of the suggestion made to me by Charles Misner to use, instead, an arbitrarily chosen timelike vector field.

CORRECTIONS

Page 58, column 2, line 12

"of a non-compact initial Cauchy ... "

Page 59, column 1, line 2

" ... condition), C_3 being a non-compact initial boundary to $\mathscr{H}^4_+$. (iv) ... "

Page 59, column 1, line 22

" ... Then $(A^{\frac{1}{2}})_{;\mu}\ell^{\mu} = -A^{\frac{1}{2}}\rho$; $-(A^{\frac{1}{2}}\rho)_{;\mu}\ell^{\mu} = \ldots$ "

GRAVITATIONAL COLLAPSE AND SPACE-TIME SINGULARITIES

Roger Penrose
Department of Mathematics, Birkbeck College, London, England
(Received 18 December 1964)

The discovery of the quasistellar radio sources has stimulated renewed interest in the question of gravitational collapse. It has been suggested by some authors[1] that the enormous amounts of energy that these objects apparently emit may result from the collapse of a mass of the order of $(10^6\text{-}10^8)M_\odot$ to the neighborhood of its Schwarzschild radius, accompanied by a violent release of energy, possibly in the form of gravitational radiation. The detailed mathematical discussion of such situations is difficult since the full complexity of general relativity is required. Consequently, most exact calculations concerned with the implications of gravitational collapse have employed the simplifying assumption of spherical symmetry. Unfortunately, this precludes any detailed discussion of gravitational radiation—which requires at least a quadripole structure.

The general situation with regard to a spherically symmetrical body is well known.[2] For a sufficiently great mass, there is no final equilibrium state. When sufficient thermal energy has been radiated away, the body contracts and continues to contract until a physical singularity is encountered at $r=0$. As measured by local comoving observers, the body passes within its Schwarzschild radius $r=2m$. (The densities at which this happens need not be enormously high if the total mass is large enough.) To an outside observer the contraction to $r=2m$ appears to take an infinite time. Nevertheless, the existence of a singularity presents a serious problem for any complete discussion of the physics of the interior region.

The question has been raised as to whether this singularity is, in fact, simply a property of the high symmetry assumed. The matter collapses radially inwards to the single point at the center, so that a resulting space-time catastrophe there is perhaps not surprising. Could not the presence of perturbations which destroy the spherical symmetry alter the situation drastically? The recent rotating solution of Kerr[3] also possesses a physical singularity, but since a high degree of symmetry is still present (and the solution is algebraically special), it might again be argued that this is not representative of the general situation.[4] Collapse without assumptions of symmetry[5] will be discussed here.

Consider the time development of a Cauchy hypersurface C^3 representing an initial matter distribution. We may assume Einstein's field equations and suitable equations of state governing the matter. In fact, the only assumption made here about these equations of state will be the non-negative definiteness of Einstein's energy expression (with or without cosmological term). Suppose this matter distribution undergoes gravitational collapse in a way which, at first, qualitatively resembles the spherically symmetrical case. It will be shown that, after a certain critical condition has been fulfilled, deviations from spherical symmetry cannot prevent space-time singularities from arising. If, as seems justifiable, actual physical singularities in space-time are not to be permitted to occur, the conclusion would appear inescapable that inside such a collapsing object at least one of the following holds: (a) Negative local energy occurs.[6] (b) Einstein's equations are violated. (c) The space-time manifold is incomplete.[7] (d) The concept of space-time loses its meaning at very high curvatures—possibly because of quantum phenomena.[2] In fact (a), (b), (c), (d) are somewhat interrelated, the distinction being partly one of attitude of mind.

Before examining the asymmetrical case, consider a spherically symmetrical matter distribution of finite radius in C^3 which collapses symmetrically. The empty region surrounding the matter will, in this case, be a Schwarzschild field, and we can conveniently use the metric $ds^2 = -2drdv + dv^2(1-2m/r) - r^2(d\theta^2 + \sin^2\theta d\varphi^2)$, with an advanced time parameter v to describe it.[8] The situation is depicted in Fig. 1. Note that an exterior observer will always see matter outside $r = 2m$, the collapse through $r = 2m$ to the singularity at $r = 0$ being invisible to him.

After the matter has contracted within $r = 2m$, a spacelike sphere S^2 ($t = \text{const}$, $2m > r = \text{const}$) can be found in the empty region surrounding the matter. This sphere is an example of what will be called here a trapped surface—defined generally as a closed, spacelike, two-surface T^2 with the property that the two systems of null geodesics which meet T^2 orthogonally converge locally in future directions at T^2. Clearly trapped surfaces will still exist if the matter region has no sharp boundary or if spherical symmetry is dropped, provided that the deviations from the above situation are not too great. Indeed, the Kerr solutions with $m > a$ (angular momentum ma) all possess trapped surfaces, whereas those for which $m \leq a$ do not.[9] The argument will be to show that the existence of a trapped surface implies—irrespective of symmetry—that singularities necessarily develop.

The existence of a singularity can never be inferred, however, without an assumption such as completeness for the manifold under consideration. It will be necessary, here, to suppose that the manifold M_+^4, which is the future time development of an initial Cauchy hypersurface C^3 (past boundary of the M_+^4 region), is in fact null complete into the future. The various assumptions are, more precisely, as follows: (i) M_+^4 is a nonsingular (+−−−) Riemannian manifold for which the null half-cones form two separate systems ("past" and "future"). (ii) Every null geodesic in M_+^4 can be extended into the future to arbitrarily large affine parameter values (null completeness). (iii) Every timelike or null geodesic in M_+^4 can be extended

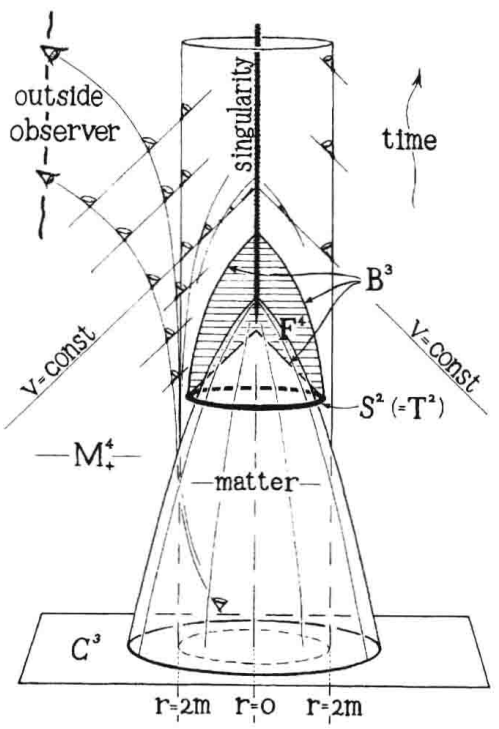


FIG. 1. Spherically symmetrical collapse (one space dimension surpressed). The diagram essentially also serves for the discussion of the asymmetrical case.

into the past until it meets C_3 (Cauchy hypersurface condition). (iv) At every point of M_+^4, all timelike vectors t^μ satisfy $(-R_{\mu\nu}+\frac{1}{2}Rg_{\mu\nu}-\lambda g_{\mu\nu})t^\mu t^\nu \geq 0$ (non-negativeness of local energy). (v) There exists a trapped surface T^2 in M_+^4. It will be shown here, in outline, that (i), $\cdots$, (v) are together inconsistent.

Let F^4 be the set of points in M_+^4 which can be connected to T^2 by a smooth timelike curve leading into the future from T^2. Let B^3 be the boundary of F^4. Local considerations show that B^3 is null where it is nonsingular, being generated by the null geodesic segments which meet T^2 orthogonally at a past endpoint and have a future endpoint if this is a singularity (on a caustic or crossing region) of B^3. Let l^μ (subject to $l^\mu{}_{;\nu}l^\nu=0$), ρ $(=-\frac{1}{2}l^\mu{}_{;\mu})$, and $|\sigma|$ $\{=[\frac{1}{2}l_{(\mu;\nu)}l^{\mu;\nu}-\frac{1}{4}(l^\mu{}_{;\mu})^2]^{1/2}\}$ be, respectively, a future-pointing tangent vector, the convergence, and the shear for these null geodesics,[10] and let A be a corresponding infinitesimal area of cross section of B^3. Then $[(A^{1/2})_{;\mu}l^\mu]_{;\nu}l^\nu = -(A^{1/2}\rho)_{;\mu}l^\mu = -A^{1/2}(|\sigma|^2+\Phi) \leq 0$ where $\Phi = -\frac{1}{2}R_{\mu\nu}l^\mu l^\nu$ $[\geq 0$ by (iv)$]$. Since T^2 is trapped, $\rho>0$ at T^2, whence A becomes zero at a finite affine distance to the future of T^2 on each null geodesic. Each geodesic thus encounters a caustic. Hence B^3 is compact (closed), being generated by a compact system of finite segments. We may approximate B^3 arbitrarily closely by a smooth, closed, spacelike hypersurface B^{3*}. Let K^4 denote the set of pairs (P,s) with $P\in B^{3*}$ and $0\leq s\leq 1$. Define a continuous map μ: $K^4 \to M_+^4$ where, for fixed P, $\mu\{(P,s)\}$ is the past geodesic segment normal to B^{3*} at $P=\mu\{(P,1)\}$ and meeting C^3 [as it must, by (iii)] in the point $\mu\{(P,0)\}$. At each point Q of $\mu\{K^4\}$, we can define the degree $d(Q)$ of μ to be the number of points of K^4 which map to Q (correctly counted). Over any region not containing the image of a boundary point of K^4, $d(Q)$ will be constant. Near B^{3*}, μ is 1-1, so $d(Q)=1$. It follows that $d(Q)=1$ near C^3 also, whence the degree of the map $B^{3*}\to C^3$ induced by μ when $s=0$ must also be unity. The impossibility of this follows from the noncompactness of C^3.

Full details of this and other related results will be given elsewhere.

---

[1]F. Hoyle and W. A. Fowler, Monthly Notices Roy. Astron. Soc. 125, 169 (1963); F. Hoyle, W. A. Fowler, G. R. Burbidge, and E. M. Burbidge, Astrophys. J. 139, 909 (1964); W. A. Fowler, Rev. Mod. Phys. 36, 545 (1964); Ya. B. Zel'dovich and I. D. Novikov, Dokl. Akad. Nauk SSSR 155, 1033 (1964) [translation: Soviet Phys.—Doklady 9, 246 (1964)]; I. S. Shklovskiĭ and N. S. Kardashev, Dokl. Akad. Nauk SSSR 155, 1039 (1964) [translation: Soviet Phys.—Doklady 9, 252 (1964)]; Ya. B. Zel'dovich and M. A. Podurets, Dokl. Akad. Nauk SSSR 156, 57 (1964) [translation: Soviet Phys.—Doklady 9, 373 (1964)]. Also various articles in the Proceedings of the 1963 Dallas Conference on Gravitational Collapse (University of Chicago Press, Chicago, Illinois, 1964).

[2]J. R. Oppenheimer and H. Snyder, Phys. Rev. 56, 455 (1939). See also J. A. Wheeler, in Relativity, Groups and Topology, edited by C. deWitt and B. deWitt (Gordon and Breach Publishers, Inc., New York, 1964); and reference 1.

[3]R. P. Kerr, Phys. Rev. Letters 11, 237 (1963).

[4]See also E. M. Lifshitz and I. M. Khalatnikov, Advan. Phys. 12, 185 (1963).

[5]See also P. G. Bergmann, Phys. Rev. Letters 12, 139 (1964).

[6]The negative energy of a "C field" may be invoked to avoid singularities: F. Hoyle and J. V. Narlikar, Proc. Roy. Soc. (London) A278, 465 (1964). However, it is difficult to see how even the presence of negative energy could lead to an effective "bounce" if local causality is to be maintained.

[7]The "I'm all right, Jack" philosophy with regard to the singularities would be included under this heading!

[8]D. Finkelstein, Phys. Rev. 110, 965 (1959).

[9]The case $m<a$ is interesting in that here a singularity is "visible" to an outside observer. Whether or not "visible" singularities inevitably arise under appropriate circumstances is an intriguing question not covered by the present discussion.

[10]For the notation, etc., see E. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962).

21
A REMARKABLE PROPERTY OF PLANE WAVES IN GENERAL RELATIVITY

Rev. Mod. Phys. **37(1)** (1965), 215–220

The singularity theorems initiated by the result of Chapter 20 stimulated an interest in the notion of global Cauchy hypersurfaces. The present paper points out that one of the most familiar solutions of the Einstein equations, namely a plane wave (now commonly referred to as a pp-wave), rather surprisingly, does not possess such a hypersurface (a fact that underlies a misconception that, at one time, led Einstein and Rosen to believe that such waves did not exist).

A remarkable property of plane waves in general relativity. *Rev. Mod. Phys.* **37(1)** (1965), 215–220

A Remarkable Property of Plane Waves in General Relativity*

ROGER PENROSE
The University of Texas, Austin, Texas†

The complete general-relativistic plane-wave space-times (gravitational, electromagnetic, or both) are examined in relation to the focusing effect they exert on null cones. The following remarkable property is then obtained. No spacelike hypersurface exists in the space-time which is adequate for the global specification of Cauchy data. As a consequence, it is not possible to imbed a plane wave globally in any hyperbolic normal pseudo-Euclidean space.

1. INTRODUCTION

Recent successes in the application of symmetry groups to strong interaction physics[1] have prompted some authors to suggest[2] that these and other new symmetries might possibly arise in some way out of space–time geometry, or from some extension of the space–time concept. Since they commute with the rotational symmetries which generate spin, the new symmetries would, according to this view, be associated with some space "orthogonal" to the four-dimensional continuum of which we are normally aware. Thus, the idea has been revived, of an *imbedding space*[3] which could possibly house, in addition to the space–time manifold itself, the extra perpendicular dimensions required for the new symmetries. The number (and signature) of these new dimensions is to be determined, in this approach, by the condition that the imbedding space be minimal and pseudo-Euclidean. It is thus of some relevance to try to determine the number and nature of the extra dimensions required, for the isometric imbedding of a general, physically interesting, general-relativistic space–time.

The problem divides naturally into two parts, that of local imbedding and that of global imbedding. It is known,[4] however, that ten dimensions—with any signature[5] -8, -6, $\cdots$, $+4$—are adequate for the *local* isometric imbedding of analytic space–time (signature -2). Furthermore, it would seem that ten dimensions are also locally *necessary* for most general-relativistic manifolds. For even if empty space–times can be always imbedded locally in fewer than ten dimensions (as seems unlikely!), it should be recalled that, in the presence of matter, Einstein's field equations amount, effectively, merely to *inequalities* on the curvature of space–time. That is, unless detailed equations for all matter fields present are assumed (and these would be exceedingly complicated when written solely in terms of the energy tensor), we are left with inequalities on the Ricci tensor which state (presumably) the positive definiteness of energy and, say, the nonnegativeness of the trace of the energy tensor. Since ten dimensions are minimal for the local isometric imbedding of a *general* 4-manifold, it seems fair to assume that ten (with a variety of possible signatures) is also correct for the local isometric imbedding of relativistic manifolds in general. The exceptions which can be locally imbedded in *fewer* than ten dimensions would from a class "of measure zero" for which special equations or symmetries, etc. might be present—although, of course, these exceptions include a large number of the known relativistic models.[6]

The problem of global imbedding[6] is much more difficult in general. It is not even known, at present, whether for a given indefinite Riemannian manifold, a global isometric imbedding into *some* pseudo-Euclidean space is always possible.[7] Furthermore, unlike the case of local imbeddings the choice of *signature* for the imbedding space seems to be important in the global case.[8] In fact, it turns out that there are *physically interesting space-times*—e.g., the plane-wave metrics—which cannot be globally isometrically imbedded in any pseudo-Euclidean n space of signature $2-n$ (hyperbolic normal signature). There are also many "unphysical" space–time models with the above property, e.g., any hyperbolic normal 4-space which contains a

* Work supported in part by the U.S. Air Force Office of Scientific Research under contract no AF-AFOSR-454-63. The larger portion of this research, however, was carried out while the author was at Princeton and Syracuse Universities in 1960–61 visiting under a NATO Fellowship administered by the DSIR in London.

† Present address: Mathematics Department, Birkbeck College, London, England.

[1] See, for example, A. Salam, Rapporteur's Report on Symmetry Theories, 1964 Dubna International Conference on High Energy Physics (to be published).

[2] Y. Ne'eman, Rev. Mod. Phys. **37**, 227 (1965), this issue.

[3] Earlier suggestions in the use of such imbeddings for a similar purpose were made by C. Fronsdal, Nuovo Cimento **13**, 988 (1959); D. W. Joseph, Phys. Rev. **126**, 319 (1962).

[4] A. Friedman, J. Math. Mech. **10**, 625 (1961).

[5] The number of positive eigenvalues of the metric tensor minus the number of negative eigenvalues.

[6] See, for example, C. Fronsdal, Phys. Rev. **116**, 778 (1959); J. Rosen, Rev. Mod. Phys. **37**, 204 (1965), preceding article.

[7] The case of positive definite Riemannian manifolds was settled in the affirmative by J. Nash, Ann. Math. **63**, 20 (1956).

[8] For example, while it is a classical theorem of D. Hilbert, Trans. Am. Math. Soc. **2**, 87 (1901), that the Lobatchevsky plane cannot be globally imbedded in ordinary Euclidean 3-space, it is easy to achieve a global imbedding as a suitable pseudosphere in pseudo-Euclidean 3-space. A result of this kind, when the space to be imbedded has an indefinite metric, has also been obtained by A. M. Garsia (private communication to Y. Ne'eman).

closed timelike smooth curve. (The Gödel cosmological model would, in this sense, be classed as "unphysical.")

Remarkably enough, the above result for plane waves is not a topological one in the most obvious sense, the plane waves having a *Euclidean* topology and the timelike and null lines being all *open*. Instead, the result rests essentially on the surprising fact, obtained here, that a plane wave admits no spacelike hypersurface which would be adequate for the global specification of Cauchy data.

2. HYPERBOLIC NORMAL IMBEDDINGS

Before discussing the plane-wave metrics in detail, let us consider the general question of isometrically imbedding a hyperbolic normal r-manifold $\mathfrak{M}_r$ (i.e., of signature $2-r$) in a hyperbolic normal pseudo-Euclidean $\mathcal{E}_n$. The $\mathcal{E}_n$ may be given by coordinates $x_0, x_1, \cdots, x_{n-1}$ and the metric

$$ds^2=dx_0^2-dx_1^2-dx_2^2-\cdots-dx_{n-1}^2. \qquad (2.1)$$

The manifold $\mathfrak{M}_r$ is to be a smooth r-dimensional subset of $\mathcal{E}_n$ whose induced metric given by (2.1) has signature $2-r$. This amounts to saying that the tangent r space to $\mathfrak{M}_r$ at any one of its points contains a timelike vector of $\mathcal{E}_n$. ["Timelike" means that the "squared length" of the vector, according to the metric (2.1), is positive. Similarly, "null" would mean that this "squared length" is zero, etc. The terms "past" and "future" will refer to the ordering with respect to x_0, which is thus thought of as a "time" coordinate.] At each point of $\mathfrak{M}_r$, the tangent r space contains two half-cones of nonzero null vectors—the *past* and *future* half-cones. The systems of past and future half-cones are thus disconnected from each other.

Consider, now, the section of $\mathcal{E}_n$ by some spacelike hyperplane $\mathcal{H}_{n-1}$ meeting $\mathfrak{M}_r$ given, say, by $x_0=c$ (where c is constant). The induced metric on $\mathcal{H}_{n-1}$ is negative definite, so the induced metric on the intersection $\mathcal{S}_{r-1}$, of $\mathcal{H}_{n-1}$ with $\mathfrak{M}_r$, is also negative definite. That is to say, $\mathcal{S}_{r-1}$ is a *spacelike hypersurface* in $\mathfrak{M}_r$. It must, in fact, be a *nonsingular* (i.e., smooth) hypersurface, since $\mathcal{H}_{n-1}$ can be nowhere tangent to $\mathfrak{M}_r$. Denote the parts of $\mathcal{E}_n$ and $\mathfrak{M}_r$, for which $x_0<c$ by $\mathcal{E}_n^-$ and $\mathfrak{M}_r^-$, respectively. Similarly, denote those parts for which $x_0>c$ by $\mathcal{E}_n^+$ and $\mathfrak{M}_r^+$. Note that $\mathcal{S}_{r-1}$ is the boundary between the two disconnected portions $\mathfrak{M}_r^-$ and $\mathfrak{M}_r^+$ of the manifold $\mathfrak{M}_r$.

Observe that any smooth connected timelike or null curve in $\mathcal{E}_n$ meets $\mathcal{H}_n$ in at most one point, since x_0 may be used as a parameter on the curve. Choose any point P in $\mathcal{E}_n^+$, with coordinates $(x_0, x_1, \cdots, x_{n-1})$. Let $\mathcal{C}_1$ be any smooth connected timelike curve in $\mathcal{E}_n$ whose future end point is P. Let $\mathcal{L}_1$ be the straight line segment joining P to the point $(c, x_1, \cdots, x_{n-1})$. The length $(dx_0^2-dx_1^2-\cdots-dx_{n-1}^2)^{\frac{1}{2}}$ of any element $d\mathcal{C}_1$ of $\mathcal{C}_1$ is clearly not more than the length dx_0 of the orthogonal projection $d\mathcal{L}_1$ of $d\mathcal{C}_1$ on $\mathcal{L}_1$. Hence the total length of $\mathcal{C}_1$ is at most x_0-c. Thus, the timelike curves in $\mathcal{E}_n$ which extend into the past from P have bounded total length. A corresponding statement holds if P is in $\mathcal{E}_n^-$.

We may specialize these statements to the case when P and $\mathcal{C}_1$ lie on $\mathfrak{M}_r$ and collect together the following necessary conditions for the imbeddability of $\mathfrak{M}_r$ in $\mathcal{E}_n$.

(2.2) *The null half-cones of $\mathfrak{M}_r$ form two systems disconnected from one another—we may call these the systems of "past" and "future" half-cones of $\mathfrak{M}_r$. To each point of $\mathfrak{M}_r$ corresponds one past and one future half-cone.*

(2.3) *There exists a nonsingular spacelike hypersurface $\mathcal{S}_{r-1}$ in $\mathfrak{M}_r$, the removal of which separates $\mathfrak{M}_r$ into two disconnected portions: $\mathfrak{M}_r^+$, which lies on the future side of $\mathcal{S}_{r-1}$, and $\mathfrak{M}_r^-$, which lies on the past side of $\mathcal{S}_{r-1}$, and furthermore:*

(2.4) *every smooth connected timelike or null curve in $\mathfrak{M}_r$ meets $\mathcal{S}_{r-1}$ in at most one point;*

(2.5) *the smooth connected timelike curves in $\mathfrak{M}_r^+$ [resp. $\mathfrak{M}_r^-$], whose future [resp. past] end points are a given fixed point P in $\mathfrak{M}_r^+$ [resp. $\mathfrak{M}_r^-$], have a fixed upper bound to their lengths.*

The fact, physically desirable for a space–time, that $\mathfrak{M}_r$ cannot contain a *closed* smooth timelike curve is ensured by (2.3) and (2.5). For if such a curve met $\mathcal{S}_{r-1}$, this would contradict (2.3), whereas if it did not meet $\mathcal{S}_{r-1}$, neighboring open curves could be found which contradict (2.5). [However, closed smooth *null geodesics* in $\mathfrak{M}_r$ would not apparently contradict (2.2), ···, (2.5), although such curves would clearly preclude the imbedding of $\mathfrak{M}_r$ in $\mathcal{E}_n$. The condition that $\mathfrak{M}_r$ contain no closed smooth null geodesics could be added to the list (2.2), ···, (2.5) but it will not be used here.] In fact, (2.4) is an easy consequence of (2.3).

3. THE PLANE WAVES

The metric for a general gravitational–electromagnetic plane wave $\mathcal{W}_4$ can be put in the form[9]

$$ds^2=dudv+h_{ij}(u)x_ix_jdu^2-dx_i\,dx_i, \qquad (3.1)$$

[9] This metric (without electromagnetism) was given originally by H. W. Brinkman, Proc. Natl. Acad. Sci. (U.S.) **9**, 1 (1923); Math. Ann. **94**, 119 (1925) in the more general form ("plane fronted" wave) that h_{ij} is allowed to be a suitable function of the x_k as well as of u. N. Rosen [Phys. Z. Sowjetunion **12**, 366 (1937)] obtained the metric $ds^2=A(dt^2-dz^2)+B_{ij}dy_idy_j$ for plane waves (but he only considered the case when B_{ij} is diagonal). Here A and B_{ij} are to be appropriate functions of $t-z$. I. Robinson (Report to the Eddington Group, Cambridge, 1956) showed that this latter metric (with electromagnetism included) can be transformed into the form (2.1) by a coordinate transformation and that the singularities that Rosen found to be necessary with his form of metric were, in fact, spurious coordinate singularities. See also Ref. 10.

The existence of Rosen's coordinate singularities is closely related to the phenomenon that is discussed here. The time development of the hypersurface $t=0$ in the Rosen metric gives only that part of $\mathcal{W}_4$ covered nonsingularly by his coordinates, i.e., $t=0$ is not a global Cauchy hypersurface for $\mathcal{W}_4$.

where

$$h_{ii}(u) \geqslant 0. \tag{3.2}$$

The indices $i, j, \cdots$, take the two values 1, 2 and the summation convention is used throughout. All variables u, v, x_1, x_2 range from $-\infty$ to $+\infty$ and the entire manifold $\mathcal{W}_4$ is covered in a $(1-1)$ fashion by this one nonsingular coordinate patch. Without loss of generality, we may take $h_{ij}(u)$ as symmetric:

$$h_{ij} = h_{ji}. \tag{3.3}$$

The condition for a purely *gravitational* wave (vanishing Ricci tensor) is that the equality in (3.2) should hold:

$$h_{ii}(u) = 0. \tag{3.4}$$

For a purely *electromagnetic* wave (vanishing Weyl tensor) we have

$$h_{ij}(u) = h(u)\ \delta_{ij}, \tag{3.5}$$

with $h(u) \geqslant 0$. The amplitude and polarization of the gravitational part of the wave is given, in the general case, by the trace-free part of h_{ij}. This may be chosen quite arbitrarily as a function of u. The amplitude of the electromagnetic part of the wave is measured by the square root of the trace of h_{ij}, which again may be chosen arbitraily as a function of u. The polarization of the electromagnetic part of the wave may also be chosen arbitraily as a function of u but it does not contribute in any way to the curvature. (The electromagnetic field is null everywhere and independent duality rotations can be applied in each hyperplane $u = \text{const}$. The gravitational field is also null.)

There is, thus, exactly as much freedom in the construction of the full general-relativistic plane waves as is the case for their linear approximations (spin 1 and 2 zero-mass waves in Minkowski space). There is also exactly as much *symmetry*, namely a five parameter group of motions[10] in general. This group acts transitively on the ∞^5 system of null geodesics of $\mathcal{W}_4$, excepting those null geodesics parallel to the propagation world-direction:

$$u = \text{const.}, \qquad x_i = \text{const.} \tag{3.6}$$

Thus, the null geodesics, which are *not* the propagation lines (3.6), are all equivalent to each other in $\mathcal{W}_4$.

It should be emphasized also that the space–time $\mathcal{W}_4$ is *physically satisfactory from the point of view of causality*. The fact that it contains no smooth closed timelike or null curves is evident from the form of the metric. On any timelike or null curve not parallel to the propagation world-direction (3.6), we can use u as a parameter continuously increasing with time. On a null curve (geodesic) parallel to this direction, v serves as such a parameter.

A particular type of plane wave whose properties are easily studied is the "sandwich" wave.[10] In this case we suppose that the gravitational and electromagnetic amplitudes vanish outside a certain range (a, b) of values of u, but that they do not both vanish everywhere. Thus

$$h_{ij}(u) = 0 \text{ unless } a < u < b; \qquad \text{some } h_{ij}(u) \neq 0 \tag{3.7}$$

so that the space–time is flat for $u < a$ and again flat for $u > b$, but it is curved in between (see Fig. 1). For simplicity, the arguments will be given here for the case of sandwich waves only, which are "sufficiently weak" (in a sense to be defined shortly). In fact, general plane waves can also be treated by essentially the same method.

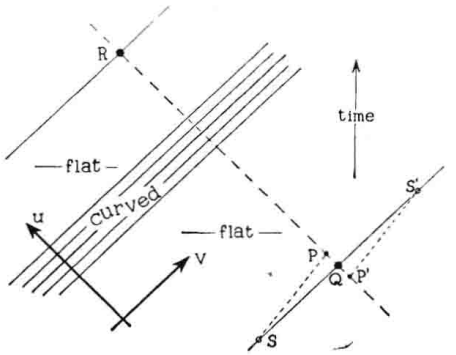


FIG. 1. The sandwich wave $\mathcal{W}_4$ and the related configuration of points mentioned in the test. The diagram essentially represents the two-dimensional section of $\mathcal{W}_4$ by $x_i = 0$.

Let Q be a point in $\mathcal{W}_4$. The *complete null cone* $\mathcal{K}_3$ of Q is the set of points lying on all the null geodesics through Q. Now choose Q with coordinates

$$x_i = 0, \qquad v = v_0, \qquad u = u_0 < a, \tag{3.8}$$

so that Q lies in one of the *flat* regions of $\mathcal{W}_4$. Near Q, the equation of $\mathcal{K}_3$ is therefore

$$(u-u_0)(v-v_0) - x_i x_i = 0 \tag{3.9}$$

which can be written

$$v = f_{ij}(u) x_i x_j + v_0 \tag{3.10}$$

where, near Q,

$$f_{ij}(u) = (u-u_0)^{-1}\ \delta_{ij}. \tag{3.11}$$

If we choose f_{ij} as a function of u so that the hypersurface (3.10) remains null (i.e., with null normal vector) also throughout the *curved* region of $\mathcal{W}_4$, then (3.10) will be the equation of $\mathcal{K}_3$. We may choose f_{ij} symmetric:

$$f_{ij} = f_{ji}. \tag{3.12}$$

The only condition to be satisfied by f_{ij} then turns out to be

$$f_{ij}' + f_{ik} f_{kl} + h_{ij} = 0, \tag{3.13}$$

[10] H. Bondi, F. A. E. Pirani, and I. Robinson, Proc. Roy. Soc. (London) **A251**, 519 (1959).

where the prime denotes differentiation with respect to u. Given the starting value (3.11), the equation (3.13) therefore extends the definition of f_{ij} uniquely into the curved region and, with (3.10), defines $\mathcal{K}_3$, until eventually f_{ij} may be come infinite.

To examine this last possibility, note that the trace of (3.13) gives

$$f_{ii}'+\tfrac{1}{2}f_{ii}f_{jj}=-\tfrac{1}{2}(f_{ik}f_{ik}\delta_{jl}\delta_{jl}-f_{ik}\delta_{ik}f_{jl}\delta_{jl})-h_{ii}\leqslant 0$$

by Schwarz' inequality and (3.2). Thus

$$\left\{\exp\left(\frac{1}{2}\int f_{ii}\,du\right)\right\}''\leqslant 0, \tag{3.14}$$

with strict inequality holding for at least *some* values of u. If we choose Q so that u_0 is very large and negative, then in the limiting case $u_0=-\infty$ we get, from (3.11), $f_{ij}=0$ for $u<a$. That is to say, $\mathcal{K}_3$ becomes the null hyperplane $v=v_0$ in the initially flat region [see (3.10)]. Thus, initially $\{\exp(\frac{1}{2}\int f_{ii}du)\}'=0$, so from (3.14) we see that $\exp(\frac{1}{2}\int f_{ii}du)$ eventually becomes zero for some finite value of u. Hence, some component of f_{ij} must become infinite[11,12] for some least value u_1 of u, where $u_1>a$ (since $f_{ij}=0$, for $u_1\leqslant a$).

Let us suppose that the amplitudes h_{ij} become zero *before* this value is reached, that is to say

$$u_1>b. \tag{3.15}$$

This is the meaning of the notion that the wave $\mathcal{W}_4$ be "sufficiently weak," as alluded to earlier. Now if we replace the value $-\infty$, for u_0, by a finite value which is very large and negative, then for values of u near a, (3.10) will be altered only slightly. Thus u_1 will change only slightly—it will remain finite and satisfy (3.15). Thus the complete null cone $\mathcal{K}_3$ of Q (for Q sufficiently far off) necessarily encounters *singularities* at the other side of the sandwich wave—where $\mathcal{W}_4$ has again become flat.

In order to examine these singularities, we consider (3.13) in the flat space region $u>b$. The equation (3.13) may then be written

$$p_{ij}'=\delta_{ij}, \tag{3.16}$$

where p_{ij} is the matrix inverse to f_{ij}:

$$p_{ij}f_{jk}=\delta_{ik}. \tag{3.17}$$

The solution of (3.16) is

$$p_{ij}(u)=u\,\delta_{ij}-q_{ij}, \tag{3.18}$$

where q_{ij} is constant and symmetric [by (3.12)]. This gives infinite values for f_{ij} whenever u is an eigenvalue of q_{ij}. Two essentially different cases can occur. Either the eigenvalues of q_{ij} are distinct or else $q_{ij}=u_1\delta_{ij}$ (q_{ij} can be diagonalized since it is symmetric). In this latter case, (3.18) gives the same form as (3.11), that is to say, the null cone $\mathcal{K}_3$ has *two* vertices, namely, Q and also the point R with coordinates

$$x_i=0,\qquad v=v_0,\qquad u=u_1. \tag{3.19}$$

The null cone of Q is thus *focused*[13] by the wave to the *single point* R (see Fig. 2). This is the case of pure *anastygmatic* focusing and it occurs with a purely electromagnetic wave. In the gravitational case, we expect to get some resultant astygmatism[14] and the eigenvalues of q_{ij} will generally be distinct. In this case the null cone $\mathcal{K}_3$ is focused to a spacelike *line* through R, as will be seen in a moment.

The null geodesics which generate $\mathcal{K}_3$ are the curves on $\mathcal{K}_3$ with null tangent vectors. Hence by (3.1), (3.13), and the derivative of (3.10), using u as a parameter, we get for each particular geodesic,

$$x_i'=f_{ij}x_j \tag{3.20}$$

with v then given by (3.10). [Note that (3.20) and (3.13) imply the relation $x_i''=-h_{ij}x_j$, which shows that h_{ij} measures the geodesic deviation of the null rays.[14]] There is an exceptional null geodesic $\mathcal{Q}_1$ through Q which is *not* parametrized by u, however, namely the one parallel to the propagation world-direction. Thus,

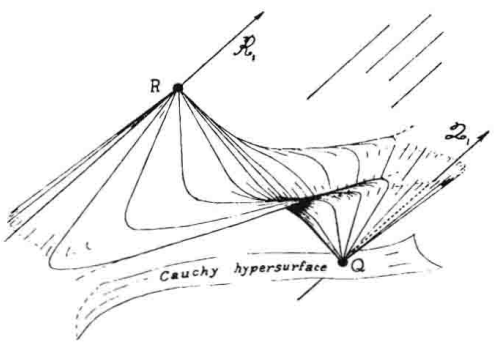


FIG. 2. The purely electromagnetic plane-wave space-times have exact analogs in two space and one time dimension. A null cone can be focused again to a second vertex. The situation is depicted above. A connected spacelike surface through Q can never meet the null line $\mathcal{R}_1$ (if the surface has no boundary).

[11] The necessary existence of singularities in the f_{ij}'s is essentially the same phenomenon as that discovered by A. Raychaudhuri, Phys. Rev. **98**, 1123 (1955); and A. Komar, Phys. Rev. **104**, 544 (1956). Here the result is applied to null geodesics rather than timelike ones. A case of this particular phenomenon was also found by I. Ozsváth and E. Schücking in *Recent Developments in General Relativity* (Pergamon Press, Ltd., London–Warsaw, 1962).

[12] This fact has relevance to the question of two colliding weak plane sandwich waves. Each wave warps the other until singularities in the wave fronts ultimately appear. This, in fact, causes the space-time to acquire genuine physical singularities in this case. The warping also produces a scattering of each wave after collision so that they cease to be sandwich waves when they separate (and they are no longer plane—although they have a two-parameter symmetry group).

[13] This focusing effect and its possible relation to energy flux is discussed in greater generality in R. Penrose, Hlavaty Festschr. (to be published).

[14] R. K. Sachs, Proc. Roy. Soc. (London) **A264**, 309 (1961), F. A. E. Pirani and A. Schild, Bull. Acad. Polon. Sci. Ser. Sci. Math. Astron. Phys. **9**, 543 (1961).

$\mathcal{Q}_1$ is given [see (3.6)] by

$$u=u_0, \qquad x_i=0. \tag{3.21}$$

When $\mathcal{K}_3$ enters the flat region $u>b$, we can solve (3.20) using (3.18) [see (3.17)] and obtain

$$x_i=(u\,\delta_{ij}-q_{ij})m_j. \tag{3.22}$$

where the m_j are constant for each geodesic. When u_1 is a repeated eigenvalue of q_{ij} (the anastygmatic case) we see that $x_i=0$ for $u=u_1$ so that *all the null geodesics through Q, with the single exception of $\mathcal{Q}_1$, pass through the point R* [given by (3.19)]. When u_1 is a simple eigenvalue of q_{ij} (the general astygmatic case), we may, for convenience, choose the original x_1 and x_2 axes in (3.1) so that q_{ij} is diagonal with $q_{11}=u_1$, $q_{12}=0$, $q_{22}=u_2(u_2\neq u_1)$. Thus, $x_1=0$ when $u=u_1$, so that *all the null geodesics through Q, with the single exception of $\mathcal{Q}_1$, pass through some point of the spacelike line*: $x_1=0$, $v=v_0$, $u=u_1$. In fact, through each point of this line there will pass the ∞^1 null geodesics given when m_2 takes a fixed value. In particular, the null geodesics through both Q and R are those given when the linear relation $m_2=0$ holds. The linearity of (3.20) then implies that the values of x_i at points near Q, for these geodesics, are also determined by a linear relation: $x_ic_i=0$ (c_i constant). That is, near Q, the null geodesics on $\mathcal{K}_3$ which eventually pass through R are the ones—except for $\mathcal{Q}_1$—lying on the intersection of $\mathcal{K}_3$ with the timelike hyperplane $x_ic_i=0$. This is a *connected* ∞^1 system of curves (the generators of an ordinary three-dimensional cone). Thus in *both* the above cases we have the property that there are null geodesics through Q, arbitrarily close to $\mathcal{Q}_1$ which pass through a *fixed* point R *not* on $\mathcal{Q}_1$.

The roles of Q and R may be interchanged. If $\mathcal{R}_1$ is the null geodesic through R parallel to the propagation direction (i.e., $\mathcal{R}_1$ has equation $u=u_1$, $v=v_0$, $x_i=0$), then $\mathcal{R}_1$ does *not* pass through Q, but neighboring null geodesics to $\mathcal{R}_1$ through R *do* pass through Q—these being the same null geodesics as in the previous case.[15] Thus, the null geodesics through *both* Q and R form either a two-dimensional or a one-dimensional system, a limiting position of the geodesics of this system giving the *pair* of null geodesics $\mathcal{Q}_1$, $\mathcal{R}_1$. Hence $\mathcal{W}_4$ has a characteristic property:

$\mathcal{W}_4$ contains a sequence of null geodesics which converges on a pair of nonintersecting null geodesics. (3.23)

The convergence in (3.23) is to be taken in the sense that any neighborhood of a point on either of the limit lines contains a point of every geodesic sufficiently far on in the sequence. There is clearly a sense in which this convergence is nonuniform. We can also allow that more than two null geodesics may emerge in the limit, although this will not in fact happen with the "sufficiently weak" waves considered here.

One consequence of (3.23) is that $\mathcal{W}_4$ contains no global Cauchy hypersurface. For a hypersurface to be completely satisfactory globally, for the specification of Cauchy data,[16] we would require that every null geodesic should intersect it exactly once. However, if each null geodesic of the sequence in (3.23) were to meet the hypersurface just once, it is clear that *at most one* of the limiting geodesics could meet it—the other would have to miss the hypersurface altogether. The situation can be made more graphic if we examine the diagram of Fig. 2. A connected spacelike surface through Q must initially lie entirely "below" (i.e., to the past of) the future null cone of Q. As this cone folds down in order to focus again at the point R, the hypersurface gets trapped beneath it (it cannot cross the null cone and remain spacelike everywhere). Thus the hypersurface can never intersect $\mathcal{R}_1$ nor any of the propagation lines beyond $\mathcal{R}_1$. Cauchy data on such a hypersurface could thus give no information for specifying amplitudes for a parallel wave which might lie beyond $\mathcal{R}_1$.

The situation of (3.23) also serves to prove the nonimbeddability of $\mathcal{W}_4$ in a hyperbolic normal $\mathcal{E}_n$, by means of a similar argument to the above. For suppose that $\mathcal{W}_4$ can be isometrically imbedded in $\mathcal{E}_n$. Then there exists a nonsingular spacelike hypersurface $\mathcal{S}_3$ in $\mathcal{W}_4$ which is in accordance with (2.3), (2.4), and (2.5) (with $\mathcal{W}_4$ for $\mathfrak{M}_r$). Thus, by (2.4), every null geodesic of the sequence in (3.23) must meet $\mathcal{S}_3$ in at most one point. Hence, $\mathcal{S}_3$ cannot meet *both* of $\mathcal{Q}_1$, $\mathcal{R}_1$. Suppose, first, that $\mathcal{S}_3$ does not meet $\mathcal{Q}_1$. Then, by (2.3) either $\mathcal{Q}_1$ lies entirely in $\mathcal{W}_4^+$ or entirely in $\mathcal{W}_4^-$. Suppose $\mathcal{Q}_1$ lies in $\mathcal{W}_4^+$. Choose a fixed point P, with $x_i=0$, $v=v_0$, $u=u_0+\epsilon$, where $a-u_0>\epsilon>0$. Then since P and $\mathcal{Q}_1$ both lie in the flat region $u<a$, there will be a straight timelike segment joining P to any point S on the half of $\mathcal{Q}_1$ which lies to the past of Q (see Fig. 1) We may specify S by $x_i=0$, $v=v_0-s$, $u=u_0$ with $s>0$. Then the Minkowski length of the segment SP will be $s\epsilon$ [see (3.1)]. This is clearly unbounded for $s\to\infty$, whereas SP remains within $\mathcal{W}_4^+$. Thus, (2.5) is contradicted. Similarly, if $\mathcal{Q}_1$ lies in $\mathcal{W}_4^-$, we choose P' with coordinates $x_i=0$, $v=v_0$, $u=u_0-\epsilon$ and S' with $x_i=0$, $v=v_0+s$, $u=u_0$ and let $s\to\infty$. Again (2.5) is contradicted. The argument is essentially the same if $\mathcal{S}_3$ does not meet $\mathcal{R}_1$.

4. CONCLUDING REMARKS

In order to assess the relevance of this result to the question of imbedding "physically sensible" spacetimes in a hyperbolic normal $\mathcal{E}_n$, we must examine how "physically sensible" the general relativistic plane waves really are. It seems fair to assume that they are

[15] Note that, somewhat remarkably, $\mathcal{K}_3$ is not a closed set. The null geodesic $\mathcal{R}_1$ is not part of $\mathcal{K}_3$ but it consists of limit points of $\mathcal{K}_3$.

[16] See, for example, Y. Bruhat in *Gravitation: An Introduction to Current Research*, edited by L. Witten (John Wiley & Sons, Inc., New York, 1963), and the references cited therein.

no *less* physical (as idealizations) than their flat space counterparts. But there are, of course, difficulties encountered with plane waves even in flat space. The total energy, for example, is infinite. Whereas the local energy flux remains meaningful for an electromagnetic wave, this is not so in the spin-2 zero-mass case. Also, the angular momentum due to the photon spin in a circularly polarized plane electromagnetic wave cannot be exhibited locally in a gauge invariant way; etc. These concepts all become meaningful only for waves (e.g., wave packets) whose amplitudes fall off suitable in *all* spacial directions so that the *total* energy–momentum and angular momentum, etc., can be defined.[17]

In the general relativistic case, the equivalent condition is usually interpreted as some form of asymptotic flatness condition for the space–time manifold—although some other assumptions are presumably appropriate on the cosmological scale. A question of some interest, therefore, is whether the somewhat strange properties of plane waves encountered here will still be present for waves which locally approximate plane waves, but for which the space–time is asymptotically flat, or asymptotically cosmological in some appropriate sense.

A geometrical approach to the global question of asymptotic flatness has been recently proposed by the author[18] which has relevance to this problem. Moreover, various asymptotically cosmological situations can also be treated by essentially the same method. A space–time is termed *asymptotically simple* if, from the point of view of its *conformal* structure, a sufficiently extensive boundary hypersurface $\mathscr{I}_3$ can be introduced "at infinity" which satisfies certain regularity conditions. For an asymptotically *flat* space–time, it turns out that $\mathscr{I}_3$ is *null*. The cases when $\mathscr{I}_3$ is *spacelike* are also of some interest cosmologically, since these cases are, in an appropriate sense, asymptotically *de Sitter* with a positive cosmological constant. The asymptotic simplicity condition then implies that in these cases the Riemann tensor falls off suitably (the "peeling off" property) for Einstein–Maxwell fields. In particular, the Weyl and Maxwell tensors must fall off as r^{-1} along every null geodesic, a property clearly not shared by the plane waves (r being a linear parameter on the null geodesic).

It is a simple matter to show, in fact, that the key property (3.23) *cannot occur* in any such asymptotically simple space–time. [If $\mathscr{I}_3$ is timelike—the asymptotically de Sitter cases with negative cosmological constant—then there appears to be no reason against (3.23) occuring. However, these cases appear to be of lesser interest cosmologically.] On the other hand, the focusing effect of a wave on null cones encountered here is a quite general phenomenon[13]—although it is peculiarly exact in the case of plane waves. In the case of a wave packet, the effect would be quite similar to that depicted in Fig. 2 for null geodesics fairly close to QR, but the extreme deflections of null geodesics through Q near $\mathscr{Q}_1$ would not take place. It is still an open question, therefore, whether global Cauchy hypersurfaces or global imbeddings in some hyperbolic normal $\mathscr{E}_n$ will always exist for wave packets.

ACKNOWLEDGMENTS

The author is grateful to I. Robinson and M. Cahen for valuable discussions.

[17] The question of including a source for the wave should also be considered to be completely physically realistic. However this seems to be not essential for the case of wave packets.

[18] R. Penrose, Phys. Rev. Letters **10**, 66 (1963); in *Relativity Groups and Topology* (the 1963 Les Houches lectures), edited by C. DeWitt and B. DeWitt (Gordon & Breach Publishers, Inc., New York, 1964). A more complete account is due to appear shortly in Proc. Roy. Soc. (London).

22
ZERO REST-MASS FIELDS INCLUDING GRAVITATION: ASYMPTOTIC BEHAVIOUR

Proc. Roy. Soc. London Ser. A **284(1397)** (1965), 159–203

This was my first thoroughly detailed account of the conformal technique whose initial introduction was described in Chapters 16 and 17. Here R.K. Sach's 'peeling-off' property of principal null directions (addressed also in Chapter 15 by different methods) is demonstrated to be a natural consequence of the regularity of conformal infinity. It may also be mentioned that the potentials introduced in Part I of this chapter—especially the discussion on p. [176] and the footnote thereon—had some seminal role in the introduction of twistor theory (see Chapter 30). The techniques that I had been developing for the Appendix also had seminal importance for the singularity theorem of Chapter 20. The Appendix argument given here is a little incomplete, and a much more elegant argument proving this result was given later by Robert Geroch (Geroch, R., 1971, Space-time structure from a global view point, In: *General Relativity and Cosmology*, Proc. of Int. Sch. in Phys. 'Enrico Fermi', Course XLVII, Ed. R.K. Sachs (Academic Press, New York), pp. 71–103).

488 Zero rest-mass fields including gravitation: Asymptotic behaviour. *Proc. Roy. Soc. London Ser. A* **284(1397)** (1965), 159–203

Zero rest-mass fields including gravitation: asymptotic behaviour

By R. Penrose†

University of Texas, Austin, Texas

(*Communicated by H. Bondi, F.R.S.—Received* 8 *June* 1964)

A zero rest-mass field of arbitrary spin s determines, at each event in space-time, a set of $2s$ principal null directions which are related to the radiative behaviour of the field. These directions exhibit the characteristic 'peeling-off' behaviour of Sachs, namely that to order r^{-k-1} ($k = 0, \ldots, 2s$), $2s-k$ of them coincide radially, r being a linear parameter in any advanced or retarded radial direction. This result is obtained in part I for fields of any spin in special relativity, by means of an inductive spinor argument which depends ultimately on the appropriate asymptotic behaviour of a very simple Hertz-type complex scalar potential. Spin ($s-\frac{1}{2}$) fields are used as potentials for spin s fields, etc. Several examples are given to illustrate this, In particular, the method is used to obtain physically sensible singularity-free waves for each spin which can be of any desired algebraic type.

In part II, a general technique is described, for discussing asymptotic properties of fields in curved space-times which is applicable to all asymptotically flat or asymptotically de Sitter space-times. This involves the introduction of 'points at infinity' in a consistent way. These points constitute a hypersurface boundary $\mathscr{I}$ to a manifold whose interior is conformally identical with the original space-time. Zero rest-mass fields exhibit an essential conformal invariance, so their behaviour at 'infinity' can be studied at this hypersurface. Continuity at $\mathscr{I}$ for the transformed field implies that the 'peeling-off' property holds. Furthermore, if the Einstein empty-space equations hold near $\mathscr{I}$ then continuity at $\mathscr{I}$ for the transformed gravitational field is a consequence. This leads to generalizations of results due to Bondi and Sachs. The case when the Einstein–Maxwell equations hold near $\mathscr{I}$ is also similarly treated here. The hypersurface $\mathscr{I}$ is space-like, time-like or null according as the cosmological constant is positive, negative or absent. The technique affords a covariant approach to the definition of radiation fields in general relativity. If $\mathscr{I}$ is not null, however, the radiation field concept emerges as necessarily origin dependent. Further applications of the technique are also indicated.

1. Introduction

The main purpose of this paper is to introduce, in some detail, a new, widely applicable technique for dealing with asymptotic questions in (special or) general relativity. This will be described at length in part II. Brief outlines of this approach have recently been given elsewhere (Penrose 1962, 1963, 1964). An additional purpose is to discuss zero rest-mass fields of arbitrary (integral or half-integral) spin and their radiation properties in general, as most of the discussion given here applies equally to any such field. The gravitational, electromagnetic and neutrino fields can thus be treated concurrently.

Zero rest-mass fields apparently exhibit a certain characteristic asymptotic behaviour (Sachs 1961, 1962; Newman & Penrose 1962; Goldberg & Kerr 1964), which Sachs has descriptively termed the *peeling-off* property. Part I is concerned primarily with an independent derivation of this important property, and some closely related questions, for fields in *flat* space-time. Then in part II the property

† Present address: Mathematics Department, Birkbeck College, London.

is derived for a class of space-times which includes all asymptotically flat or asymptotically de Sitter space-times, using the technique alluded to above. The agreement achieved lends weight to a belief in the validity of the assumptions made and of the generality of the technique.

The peeling-off property is most readily described in terms of *principal null directions* (abbreviated p.n.d.) associated with the field (Synge 1956; Ruse 1944; Debever 1959; Penrose 1959, 1960). These are a special set of $2s$ directions along the light cone, defined at each event in space-time, for any spin s zero rest-mass field. The p.n.d. provide a natural means of classifying zero rest-mass fields algebraically which appears to be of great importance in the study of their radiation properties—particularly in the case of gravitation (Pirani 1957, 1959; Robinson & Trautman 1962; Jordan, Ehlers & Sachs 1961; Goldberg & Sachs 1962, etc.). The p.n.d. are only undefined at events at which the field vanishes (in which case, in a sense, *every* null direction at the event is a p.n.d.). However, some of them may coincide under certain circumstances. If all $2s$ p.n.d. coincide, the field is called a *null* field. If r is a retarded or advanced radial parameter (e.g. an affine parameter on a null geodesic), then it turns out that the part of the field that falls off as r^{-1}—the far zone or radiation field—is effectively null (compare Trautman 1958; Arnowitt, Deser & Misner 1961). If we proceed inwards from infinity (or equivalently, if we examine the field to increasing orders of accuracy) we encounter successive zones where $r^{-2}, r^{-3}, \ldots$ terms become important in turn. The behaviour of the p.n.d. that emerges is that the field in the r^{-k-1} zone ($k = 0, 1, 2, \ldots, 2s$) has, to this degree of approximation (at least and in general) $2s-k$ coincident p.n.d. pointing radially (i.e. in the direction of the radiation). This is the peeling-off property. It was obtained first by Sachs (in the case of spins 1 and 2) in flat space by a certain multipole expansion argument (Sachs 1961) and for gravitation with asymptotic flatness (Sachs 1962) based on the work of Bondi (1960) and of Bondi, van der Burg & Metzner (1962). A special case of this behaviour was also apparent for the space-times discovered by Robinson & Trautman (1962). Later, Newman & Penrose (1962) obtained the result for gravitation based on somewhat different and slightly more general assumptions and Goldberg & Kerr (1964) obtained it for retarded electromagnetism in flat space-time with bounded sources.

The main argument of part I is an essentially simple one and rests on the fact (obtained here) that, in flat space-time, a spin $(s-\frac{1}{2})$ zero rest-mass field can be used as a potential for a spin s zero rest-mass field. This leads to an inductive proof of the peeling-off property (in flat space-time) which would seem to give greater generality than can easily be obtained using multipole expansions. It also leads to the introduction of a simple Hertz-type potential which can take the form of a single complex component satisfying the wave equation. The power and simplicity obtained owes much to a consistent use of two-component spinor techniques. The arguments will apply, in particular, to the Weyl neutrino field, the free Maxwell field and the linearized Einstein gravitational field. Some simple examples of such fields obtained in this way are also given. In particular, singularity-free wave-like solutions for each spin are exhibited for which the field can have any given algebraic type (for example, such a field can be everywhere null). The peeling-off property implies

that these fields are 'physically reasonable' (i.e. unlike plane or cylindrical waves they have finite total energy-momentum.)

The argument for curved space-times given in part II is based on the fact that the zero rest-mass field equations can, with suitable interpretations, be regarded as being *conformally invariant* (see Cunningham 1910; Bateman 1910; McLennan 1956; also Buchdahl 1959). This conformal invariance is also obtained independently here in a form suitable for the present applications. In the study of zero rest-mass fields, therefore, the space-time may be considered from the point of view of its *conformal* rather than its metric structure. This has the advantage that *infinity* can be treated as an ordinary hypersurface boundary to the space-time. Asymptotic behaviour is then, effectively, behaviour in the neighbourhood of this hypersurface. A fictitious metric, conformal to the physical metric, is introduced for which physical infinity becomes a finite hypersurface. Calculations are then carried out with respect to this new metric. Thus, asymptotic behaviour for zero rest-mass fields can be studied by *local* techniques. The arguments apply both in special and in general relativity and to interacting or non-interacting fields.

The peeling-off property of the p.n.d. is shown to be equivalent to a requirement that the field should be effectively finite (and continuous) on the hypersurface representing infinity. It is also shown that this finiteness requirement can, in many cases, be deduced (the relevant field equations being assumed) from more primitive requirements concerning the asymptotic nature of the space-time itself—requirements which are satisfied if the space-time is asymptotically flat or asymptotically de Sitter. Precise geometrical definitions of asymptotic conditions to be imposed on such space-times are given here. The asymptotically de Sitter cases occur when the hypersurface is space-like (or time-like); for the asymptotically flat cases it is null. For the cases when the hypersurface is null, a proof is outlined in an appendix that its topology is, of necessity, the same as in the case of Minkowski space-time. A remarkably close agreement is then achieved, between the present geometrical approach to asymptotic flatness and that due to Bondi and Sachs, which used a special choice of co-ordinate system.

The technique also affords a covariant approach to the question of radiation fields in general relativity, which could prove valuable for the quantization programme. Outgoing and incoming fields can be considered simultaneously.† A curious but important feature arising in the case of the asymptotically de Sitter space-times is an essential origin-dependence of the radiation field concept. These are the cases for which a cosmological term is present in Einstein's equations.

It is intended that the ideas introduced here should provide a convenient mathematical framework for the further investigation of topics in which asymptotic questions are important. For example, since energy-momentum cannot be completely localized in general relativity, it is perhaps, best treated from an asymptotic

† Retarded co-ordinate systems of the Bondi type are probably not well suited to the treatment of *incoming* radiation. However, according to the analysis given here, the presence of some incoming radiation would not necessarily be *excluded* by Bondi's (1960, *et al.* 1962) co-ordinate assumptions as appears to have been suggested. Most of Bondi's 'non-radiative' motions would seem, therefore, to occur in the presence of *incoming* radiation. (A mass increase would not show up when measured at retarded times.)

point of view. A discussion of this and some other matters, on the basis of the technique described here, will be given later. Partial accounts have already appeared (Penrose 1963, 1964).

I. ZERO REST-MASS FIELDS AND THEIR PROPERTIES

2. Zero rest-mass fields

The p.n.d. and matters relating to them are most conveniently studied within the two-component spinor formalism (van der Waerden 1929; Laporte & Ulhenbeck 1931; Infield & van der Waerden 1933; Ruse 1937; Penrose 1960). A translation to tensor form is then not hard to achieve when the spin is integral. Latin capitals are used here for spinor indices and take the values 0, 1, *primed* indices referring to the complex conjugate spin-space. Greek letters are used for tensor indices and range over 0, 1, 2, 3. The metric tensor $g_{\mu\nu}$ has signature $(+---)$. The translation from tensors to spinors is achieved by means of a quantity $\sigma^{\mu}_{AB'}$ which is Hermitian for each value of μ and which satisfies

$$g_{\mu\nu}\sigma^{\mu}_{AB'}\sigma^{\nu}_{CD'} = \epsilon_{AC}\epsilon_{B'D'}, \tag{2·1}$$

where ϵ_{AB}, $\epsilon_{A'B'}$ together with ϵ^{AB}, $\epsilon^{A'B'}$ are the skew-symmetric Levi-Civita symbols (that is, they take the value 0, 1, -1, 0 according as AB is 00, 01, 10, or 11, respectively). The ϵ's are used for raising or lowering spinor indices, e.g.

$$\xi^{A} = \epsilon^{AB}\xi_{B}, \quad \xi_{B} = \xi^{A}\epsilon_{AB}, \tag{2·2}$$

and similarly for primed indices. The summation convention is used throughout this paper.

The symbol ∇_{μ}, or equivalently $\nabla_{AB'}$ related to it by

$$\nabla_{AB'} = \sigma^{\mu}_{AB'}\nabla_{\mu},$$

will be used here for *covariant* derivative. Covariant derivative of spinor quantities is uniquely determined by that for tensor quantities by the additional requirement that $\sigma^{\mu}_{AB'}$ and ϵ_{AB} be covariantly constant:

$$\nabla_{\mu}\epsilon_{AB} = 0, \quad \nabla_{\nu}\sigma^{\mu}_{AB'} = 0.$$

In *flat* space-time Minkowski co-ordinates x^{μ} may be chosen and $\sigma^{\mu}_{AB'}$ may be given constant components (say $2^{-\frac{1}{2}}\times$ the Pauli matrices and the unit matrix). Then we have

$$\nabla_{AB'} \equiv \partial/\partial x^{AB'}, \tag{2·3}$$

where

$$\sigma^{\mu}_{AB'}x^{AB'} = x^{\mu}. \tag{2·4}$$

To describe a zero rest-mass field of spins s (integral or half-integral) in flat space-time there are many possible formalisms (see for example, Dirac 1936; Fierz 1940; Fierz & Pauli 1939; Gårding 1945). For each spin, the various free-field possibilities are, however, essentially equivalent (Wigner & Bargmann 1948). The most convenient one for the present purposes is achieved by taking a totally symmetric spinor $\phi_{AB\ldots K}$ with $2s$ indices to represent the field. If $s > 0$, the free-field equations are to be

$$\nabla^{AM'}\phi_{AB\ldots K} = 0 \tag{2·5}$$

or equivalently (see (2·2))

$$\nabla_{YZ'}\phi_{AB\ldots K} = \nabla_{AZ'}\phi_{YB\ldots K}. \tag{2·6}$$

No gauge transformations are involved here since these emerge only in connexion with *potentials* (see end of §4). Equation (2·5) (or (2·6)) carries over immediately to curved space in an unambiguous way with ∇ denoting covariant derivative. (One objection to equation (2·5) would appear to have been concerned with problems connected with its behaviour under space-reflexions. However, since the discovery of parity non-conservation in nature, this objection, where relevant, can no longer be regarded as significant.)

If s is integral, then $\phi_{AB\ldots K}$ has an even number of indices. We can then introduce $2s$ primed indices by multiplying by s primed ϵ's and so a translation to tensor form can be achieved by using σ's. The real part of (twice) this resulting tensor is another way of representing the field ($s > 0$). If $s = 1$ we get a skew-symmetric $F_{\mu\nu}$ satisfying Maxwell's free-field equations (Laporte & Uhlenbeck 1931) where

$$F_{\mu\nu}\sigma^{\mu}_{AP'}\sigma^{\nu}_{BQ'} = \phi_{AB}\epsilon_{P'Q'} + \epsilon_{AB}\overline{\phi}_{P'Q'}. \tag{2·7}$$

(The bar denotes complex conjugate. Unprimed indices become primed and primed indices become unprimed under complex conjugation.) If $s = 2$ we get a $K_{\mu\nu\rho\tau}$ with the same symmetries as a Riemann tensor (Witten 1959) which satisfies Einstein's equation $K^{\mu}{}_{\nu\mu\tau} = 0$ with

$$K_{\mu\nu\rho\tau}\sigma^{\mu}_{AP'}\sigma^{\nu}_{BQ'}\sigma^{\rho}_{CR'}\sigma^{\tau}_{DS'} = \phi_{ABCD}\epsilon_{P'Q'}\epsilon_{R'S'} + \epsilon_{AB}\epsilon_{CD}\overline{\phi}_{P'Q'R'S'}. \tag{2·8}$$

The differential equation (2·5) can be written† (Penrose 1960)

$$\nabla_{[\tau}K_{\mu\nu]\rho\sigma} = 0, \tag{2·9}$$

which is formally identical with the Bianchi identity. (Square brackets denote skew-symmetrization.) Thus, in particular, in curved empty space-time without cosmological term, the *Riemann tensor* describes a spin 2 zero rest-mass field. The presence of a cosmological constant does not essentially change the situation if the Weyl conformal tensor now represents the field (see §10).

In flat-space-time, applying the operator $\nabla_{NM'}$ to (2·5) leads to

$$\Box\phi_{AB\ldots K} = 0, \tag{2·10}$$

where the D'Alembertian operator $\Box$ may be defined by

$$\Box \equiv \nabla_{AM'}\nabla^{AM'} \equiv \nabla_{\mu}\nabla^{\mu},$$

since then

$$\nabla_{YM'}\nabla^{AM'} \equiv \tfrac{1}{2}\delta^{A}_{Y}\Box. \tag{2·11}$$

However, in curved space-time, if we wish to retain (2·10), we must re-define $\Box$ by the introduction of terms involving the curvature (cf. Penrose 1960). Similarly, for the spin zero wave equation in curved space-time we should, for consistency, take

$$\Box\phi \equiv \{\nabla_{\mu}\nabla^{\mu} + \tfrac{1}{6}R\}\phi = 0, \tag{2·12}$$

where R is the scalar curvature, there being no linear equation like (2·5) for this case.

† Unlike the situation for the spin 1 case, (2·9) is *equivalent* to $\nabla^{\mu}K_{\mu\nu\rho\tau} = 0$.

3. The principal null directions

To obtain the p.n.d. for general spin, consider the expression

$$\phi_{AB\ldots K}\xi^A\xi^B\ldots\xi^K, \tag{3·1}$$

where, for simplicity, $\xi^A = (1, \xi)$. Then (3·1) is just a polynomial of degree (at most) $2s$ in ξ with complex coefficients and so it can be completely factorized into $2s$ linear factors:

$$\phi_{AB\ldots K}\xi^A\xi^B\ldots\xi^K \equiv (\alpha_A\xi^A)(\beta_B\xi^B)\ldots(\kappa_K\xi^K). \tag{3·2}$$

(If the degree of (3·1) falls short of $2s$ by an amount q, then q of $\alpha_A, \beta_B, \ldots, \kappa_K$ must have their second component zero and the corresponding factors will simply be constants.) Equating coefficients in (3·2) gives

$$\phi_{AB\ldots K} = \alpha_{(A}\beta_B\ldots\kappa_{K)}, \tag{3·3}$$

where round brackets around indices indicate that the symmetrical part is to be taken. The decomposition (3·3) is *unique* (if $\phi_{A\ldots K} \neq 0$) except that the ordering of $\alpha_A, \beta_A, \ldots, \kappa_A$ is clearly irrelevant and each of them is determined only to within multiplication by a non-zero complex number. (This uniqueness follows from the uniqueness of (3·2).) Now, each of $\alpha_A, \ldots, \kappa_A$ determines a null vector as follows:

$$\sigma_\mu^{AB'}\alpha_A\bar{\alpha}_{B'}, \ldots, \sigma_\mu^{AB'}\kappa_A\bar{\kappa}_{B'}.$$

These null vectors are determined up to proportionality by $\phi_{A\ldots K}$ and so the corresponding null directions—the *principal null directions*—are uniquely determined at each event.

Observe that ξ_A corresponds to a p.n.d. if and only if the expression (3·1) vanishes, since the right-hand side of (3·2) vanishes if and only if ξ_A is proportional to one of $\alpha_A, \ldots, \kappa_A$. (We have $\xi_A\xi^A = 0$.) Of course, it need not be assumed here that ξ_A is of the form $(1, \xi)$, so long as it does not vanish. More generally we observe from (3·3) that a necessary and sufficient condition for at least j p.n.d. to coincide along the direction corresponding to ξ_A is

$$\phi_{A\ldots FG\ldots K}\xi^G\ldots\xi^K = 0, \tag{3·4}$$

where ξ appears $2s-j+1$ times†.

The tensor translation of (3·1) leads to the following respective conditions on a null vector x^μ that it should correspond to a p.n.d.:

$$F_{\mu[\nu}x_{\rho]}x^\mu = 0, \quad x_{[\sigma}K_{\lambda]\mu\nu[\rho}x_{\tau]}x^\mu x^\nu = 0$$

in the case of spin 1 and 2 (Debever 1959; Sachs 1961).

† Note that (3·4) implies $\phi_{A\ldots GH\ldots K}\xi^H\ldots\xi^K = \mu\xi_A\ldots\xi_G$ where $2s-j$ ξ's appear on the left and j ξ's on the right. This fact can be used to prove very simply the general spinor version of a theorem due to Robinson, Bel, Sachs and others. A field is called *algebraically special* if at least two of the p.n.d. coincide at every event. Applying $\nabla^{AP'}$ to the appropriate expression (3·4) and using (2·5) gives, via the above, $\xi^A\xi^B\nabla_{AP'}\xi_B = 0$. This implies first that $\xi^B(\xi^A\bar{\xi}^{P'}\nabla_{AP'})\xi_B = 0$, the condition for the multiple p.n.d. to be tangent to a congruence of null geodesics, and secondly that the shear of this congruence should vanish (see Newman & Penrose 1962; Jordan *et al.* 1961).

4. Zero rest-mass potentials

I shall next show how all the solutions of the spin s zero rest-mass equation (2·5) in *flat space-time* can be obtained in a very simple way from the solutions for spin $(s-\frac{1}{2})$. Hence, repeating this $2s$ times, it will follow that the general solution for spin s is simply expressible in terms of the general solution of the wave equation (i.e. spin zero).

Let $\phi_{A\ldots K}$ be symmetric with $2s$ indices and satisfy the spin s zero rest-mass equation (2·5). Let $\mu^{M'}$ be a spinor which is chosen to be *constant* throughout Minkowski space-time. Then it will emerge that, locally at least, we can find a symmetric $\psi_{B\ldots K}$ with $2s-1$ indices satisfying the spin $(s-\frac{1}{2})$ equation such that

$$\phi_{AB\ldots K} = \mu^{M'}\nabla_{AM'}\psi_{B\ldots K} \tag{4·1}$$

and continuing the process leads to a 'Hertz type' complex potential χ:

$$\phi_{A\ldots K} = \mu^{M'}\ldots\mu^{W'}\nabla_{AM'}\ldots\nabla_{KW'}\chi, \tag{4·2}$$

or more generally,

$$\phi_{AB\ldots K} = \mu^{M'}\nu^{N'}\ldots\omega^{W'}\nabla_{AM'}\nabla_{BN'}\ldots\nabla_{KW'}\chi, \tag{4·3}$$

where χ is subject only to $\Box\chi = 0$ and where $\mu^{M'}$, $\nu^{N'}, \ldots, \omega^{W'}$ are given constant spinors (compare also Synge 1956 for the spin 1 case). It is clear from (2·6) that any $\phi_{A\ldots K}$ defined by (4·1) will be symmetric, and, from (2·11) and (2·10), that it will satisfy the required equation (2·5).

Conversely, to prove the possibility of (4·1) given $\phi_{A\ldots L}, \mu^{M'}$, let us for simplicity choose a Minkowski co-ordinate system and constant $\sigma^{\lambda}_{AB'}$ such that $\mu^{M'} = (1,0)$. Then

$$\mu^{M'}\nabla_{AM'} \equiv \nabla_{A0'} \equiv \frac{\partial}{\partial u^A}; \quad \nabla_{A1'} \equiv \frac{\partial}{\partial v^A};$$

where

$$u^A = x^{A0'}, \quad v^A = x^{A1'}$$

(see (2·3) (2·4)). With this notation we can write the free-field equation (2·6) as

$$\frac{\partial}{\partial u^Y}\phi_{AB\ldots K} - \frac{\partial}{\partial u^A}\phi_{YB\ldots K} = 0, \tag{4·4}$$

$$\frac{\partial}{\partial v^Y}\phi_{AB\ldots K} - \frac{\partial}{\partial v^A}\phi_{YB\ldots K} = 0. \tag{4·5}$$

The *vanishing curl* (4·4) implies that $\phi_{AB\ldots K}$ is a *gradient* in the two variables u^A, as required for (4·1):

$$\phi_{AB\ldots K} = \frac{\partial}{\partial u^A}\psi_{B\ldots K}, \tag{4·6}$$

at least locally. This point will be elaborated shortly. Clearly $\psi_{B\ldots K}$ can be symmetrized without affecting (4·6) because of the symmetry of $\phi_{AB\ldots K}$. This symmetry also implies

$$\frac{\partial}{\partial u^A}\psi_{BC\ldots K} - \frac{\partial}{\partial u^B}\psi_{AC\ldots K} = 0, \tag{4·7}$$

which is the half corresponding to (4·4) of the required spin $(s-\frac{1}{2})$ free-field equation. It remains to be shown that the remaining half

$$\frac{\partial}{\partial v^Y}\psi_{AC\ldots K} - \frac{\partial}{\partial v^A}\psi_{YC\ldots K} = 0 \tag{4·8}$$

corresponding to (4·5) can be satisfied without affecting (4·6). To show this, observe that (4·5), (4·6) and the symmetry of $\phi_{A\ldots K}$ imply

$$\frac{\partial}{\partial u^B}\left\{\frac{\partial}{\partial v^Y}\psi_{AC\ldots K}-\frac{\partial}{\partial v^A}\psi_{YC\ldots K}\right\}=0.$$

This at least tells us that the left-hand side $\theta_{YAC\ldots K}$ of (4·8) is a function of the two variables v^0, v^1 only. Thus, if we subtract, say $\int\theta_{10C\ldots K}dv^1$ from $\psi_{0C\ldots K}$, (4·6), (4·7) will be unaffected and (4·8) will be satisfied by the resulting expression.

If ψ has no indices (i.e. $s=\frac{1}{2}$), then (4·7) and (4·8) are not applicable. But from (4·6) and (4·5) we get

$$0=\left\{\frac{\partial}{\partial v^Y}\frac{\partial}{\partial u^A}-\frac{\partial}{\partial v^A}\frac{\partial}{\partial u^Y}\right\}\psi\equiv\tfrac{1}{2}\epsilon_{AY}\,\Box\psi$$

(see (2·11)) so that ψ satisfies the spin zero equation. This enables the induction to (4·2) or (4·3) to be completed, and with this notation (4·2) can be written:

$$\phi_{A\ldots K}=\frac{\partial}{\partial u^A}\cdots\frac{\partial}{\partial u^K}\chi,$$

where χ is complex, satisfying $\Box\chi=0$.

One point in the above proof strictly requires a little more consideration, namely the deduction of (4·6) from (4·4). The theorem appealed to here—that a vanishing curl on one index implies that this index can arise from a gradient operation—is normally stated for *real* variables and then this result can be extended to apply globally to any *simply-connected* region. (A simply-connected region is one in which any closed curve can be continuously contracted to a point while remaining within the region.) However, in the present application the variables are not all real so a closer examination is required. Here we have two real variables

$$u^0=x^{00'}=u,\quad v^1=x^{11'}=v,$$

say, and a pair of conjugate complex variables

$$u^1=x^{10'}=z,\quad v^0=x^{01'}=\bar{z}$$

($x^{AB'}$ being Hermitian). It is required to know the circumstances in which an equation

$$\frac{\partial a}{\partial z}=\frac{\partial b}{\partial u}$$

implies

$$a=\frac{\partial c}{\partial u},\quad b=\frac{\partial c}{\partial z}$$

for some c, where a, b and c are complex functions of u, v, z, $\bar{z}$. The proof that the implication is true *locally* is straightforward, although an integration with respect to the complex variable z is required at some stage. (The possibility of carrying out such an integration in general follows from the possibility of solving the two-dimensional Poisson equation: to solve $\partial X/\partial z=f$, put $X=\partial Y/\partial\bar{z}$ where $\nabla^2Y=4f$, with $\nabla^2\equiv 4\partial^2/\partial z\partial\bar{z}$.) Globally, the result can, in fact, be extended to any (finite) region whose sections by $v=$ constant are not only simply-connected (vanishing

first homotopy group) but have *vanishing second homotopy group* as well. (The second homotopy group of a region vanishes if every topological 2-sphere within the region can be continuously contracted within the region to a point.) This topological restriction on the region partly explains why it is not always possible to find a *global* Hertz potential for a field. For example, for a Coulomb field, the region of definition of the field excludes a timelike line. A topological 2-sphere (in v = const.) surrounding this line cannot be continuously contracted to a point without crossing the line. The question of the Coulomb field, etc., will be returned to in § 7.

It is possible to generalize the potential fields introduced here. If $\chi^{\cdots}$ is any quantity satisfying the wave equation $\Box\chi^{\cdots} = 0$ in flat space-time then (2·11) tells us that

$$\nabla_{AM'}\nabla_{BN'}\chi^{\cdots}$$

is symmetrical in M', N' and similarly it is also symmetrical in A, B. Differentiating this repeatedly and permuting the derivative symbols (permitted only in flat space-time) we see that

$$\nabla_{AM'}\ldots\nabla_{KW'}\chi^{\cdots}$$

is symmetrical in all the unprimed indices $A, \ldots, K$ and symmetrical in all the primed indices $M', \ldots, W'$. Thus, if $\chi^{M'\ldots W'}$, with $2s$ indices, is such that

$$\Box\chi^{M'\ldots W'} = 0$$

and $\phi_{A\ldots K}$ is given by

$$\phi_{A\ldots K} = \nabla_{AM'}\ldots\nabla_{KW'}\chi^{M'\ldots W'}, \tag{4·9}$$

then $\phi_{A\ldots K}$ is clearly totally symmetrical, and satisfies the spin s equation (2·6).

If $\chi^{M'N'\ldots W'}$ has the special form $\mu^{M'}\nu^{N'}\ldots\omega^{W'}\chi$, where $\mu^{M'}, \nu^{N'}, \ldots, \omega^{W'}$ are constant then evidently (4·3) emerges as a special case of (4·9). Conversely, considering (4·9) as a sum of the various expressions obtained as $M', \ldots, W'$ take on different values, we see that the expression (4·9) is simply a linear combination of expressions like (4·3). If $\chi^{M'\ldots W'}$ is of the form $\theta^{M'\ldots W'}\chi$ where $\theta^{M'\ldots W'}$ is constant and symmetric, then we can apply (3·3) to obtain a decomposition

$$\theta^{M'N'\ldots W'} = \mu^{(M'}\nu^{N'}\ldots\omega^{W')},$$

where $\mu^{M'}, \ldots, \omega^{W'}$ are constant. When substituted into (4·9), this simply gives us (4·3) again.

Observe that $\chi^{M'\ldots W'}$ can always be symmetrized without affecting (4·9). The general Hertz-type potential for a spin s zero rest-mass field can thus take the form of a *totally symmetric spinor with* $2s$ *primed indices satisfying the wave equation*. The tensor translations of these potentials are expressions with the same number of indices and symmetries as the tensor expressions for the *fields*, the equation to be satisfied being now just the wave equation. Thus for spin 1 we have a skew-symmetric $H^{\mu\nu}$ and for spin 2 a $J^{\mu\nu\rho\sigma}$, with the symmetries of a Riemann tensor, satisfying $J_{\mu}{}^{\nu\mu\sigma} = 0$. (Sachs & Bergmann 1958; Sachs 1961).

It should be noted that if $s \geqslant 1$, potentials intermediate between $\phi_{A\ldots K}$ and $\chi^{M'\ldots W'}$ can be defined (compare Fierz 1940). Thus, if

$$\psi^{M'\ldots Q'}_{F\ldots K} = \nabla_{FR'}\ldots\nabla_{KW'}\chi^{M'\ldots Q'R'\ldots W'}, \tag{4·10}$$

then it follows that $$\phi_{A\ldots EF\ldots K} = \nabla_{AM'}\ldots\nabla_{EQ'}\psi^{M'\ldots Q'}_{F\ldots K}, \tag{4·11}$$

where $\psi^{M'\ldots Q'}_{F\ldots K}$ is symmetric in each set of indices and is subject to

$$\nabla^{FL'}\psi^{M'\ldots Q'}_{FG\ldots K} = 0. \tag{4·12}$$

In fact, it is an easy consequence of the above arguments that *any* $\psi^{M'\ldots Q'}_{F\ldots K}$ which is symmetric in $F\ldots K$ (not necessarily symmetric in $M'\ldots Q'$) and which satisfies (4·12), is expressible in the form (4·10), where $\chi^{M'\ldots W'}$ (not necessarily symmetric) satisfies $\Box\chi^{M'\ldots W'} = 0$. Then $\phi_{A\ldots K}$ defined by (4·11) will be symmetric and will satisfy the spin s equation (2·5).

According to some formalisms, a spinor $\psi^{M'\ldots Q'}_{A\ldots K}$ symmetric in both sets of indices and satisfying (4·12)—or some less stringent condition which would be implied if (4·12) holds—could be regarded as representing a spin s *field* (see Fierz 1940; Gårding 1945) rather than, as here, merely a potential for a field satisfying (2·5). It should be borne in mind, however, that (4·12) is *not* irreducible under the inhomogeneous Lorentz group, having solutions which are a mixture of spins $s, s-1, \ldots, 0$ or $\frac{1}{2}$. The lower spin parts correspond to the *gauge* freedom† for $\psi^{M'\ldots Q'}_{F\ldots K}$ which does not affect the $\phi_{A\ldots K}$ defined by (4·11). Thus $\phi_{A\ldots K}$ properly represents the spin s field in a gauge-independent way (Fierz 1940), for which $\psi^{M'\ldots Q'}_{F\ldots K}$ is a gauge dependent potential. As a simple illustration, consider the case $s = 1$. Then $(\psi_{FM'} + \overline{\psi}_{M'F})\,\sigma_\mu^{FM'}$ is the usual electromagnetic vector potential satisfying the Lorenz gauge, where the field $F_{\mu\nu}$ is given by (2·7). Also the skew-Hermitian part of $\psi_{FM'}$ gives a corresponding potential for the *dual* of $F_{\mu\nu}$. Similarly, in the case $s = 2$, the Hermitian part of $\psi_{FGM'N'}$ gives the usual potential for the linearized gravitational field, etc. Potentials such as these play a necessary part in local definitions of such quantities as energy-momentum, number density (or current vector), angular momentum, scalar product, etc. (Hermitian potentials may be used when sources are present.) A detailed discussion of some of these matters, in flat and in curved space-times is left to a later paper.

5. A FLAT SPACE-TIME REFERENCE SYSTEM

In order to facilitate the arguments of §§ 6 and 7 it will be convenient to introduce a suitable reference system. Choose a space-time origin O and a time-like straight line h (of 'spatial origins') through O. Select the system of future null cones with vertices (Q) on h (although a consistent choice of past null cones would do equally well for most purposes)‡. It will be necessary to choose, for each generator of each

† It is not hard to see, for example, that if $s \geqslant 1$, $\psi^{M'}_{B\ldots K}$ is determined by $\phi_{A\ldots K}$ up to the addition of an expression $\nabla^{M'}_{K}\phi_{B\ldots J}$ where $\phi_{B\ldots J}$ is symmetric satisfying the spin $(s-1)$ free-field equation. If $s = \frac{1}{2}$, the gauge freedom is an undifferentiated (conjugate) spin $\frac{1}{2}$ field. Similarly, the extra gauge freedom for $\psi^{M'N'}_{C\ldots K}$ is just the second derivative of a spin $(s-2)$ field if $s \geqslant 2$, the first derivative of a (conjugate) spin $\frac{1}{2}$ field if $s = \frac{3}{2}$, and simply a (conjugate) spin 1 field if $s = 1$. This process continues. Thus, the extra gauge freedom for each successive potential (assumed symmetric in its primed indices) is effectively just some zero rest-mass free field of spin $\leqslant s$.

‡ Effectively, the choice of future cones means here that the properties of the outgoing rather than of the incoming radiation of a system are to be studied. This matter should become clearer from §§ 8 and 13.

cone, a linear (affine) parameter r and a spinor ξ_A corresponding to the relevant (outward) null direction. There is the freedom of choice

$$\xi_A \to \alpha\xi_A, \quad r \to ar+b$$

(α complex; a, b real) for each of these generators and the particular choice that will be made here is purely a matter of convenience.

Let P be a variable event in space-time (see figure 1) and consider the quantities $x^{AB'}$, $h^{AB'}$, ξ^A, r, t, η^A, ζ as functions of P ($h^{AB'}$ and η^A being constant) defined as

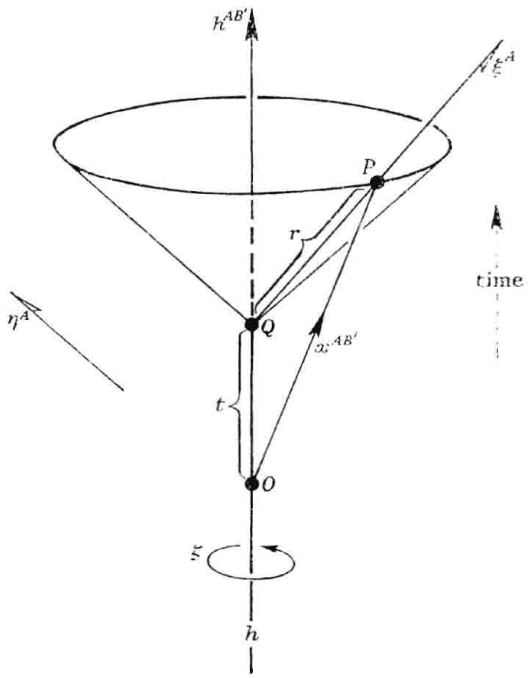


FIGURE 1. The relation between the various reference quantities of §5.

follows. The position vector relative to O of P is x^μ and $x^{AB'} = \sigma_\mu^{AB'}x^\mu$ is its spinor equivalent. Thus

$$\nabla_{CD'}x^{AB'} = \delta_C^A\delta_{D'}^{B'} \tag{5·1}$$

(see (2·3)). To represent the timelike direction of h, we may choose a Hermitian spinor $h^{AB'}$ normalized so that its determinant is unity. That is,

$$h^{AB'}h^{CD'} - h^{AD'}h^{CB'} = \epsilon^{AC}\epsilon^{B'D'}. \tag{5·2}$$

(This means that the corresponding vector $h^{AB'}\sigma^\mu_{AB'}$ has length $\sqrt{2}$ since $h^{AB'}h_{AB'} = 2$.) Let Q be the event on h such that P lies on its future null cone. Thus $\mathbf{QP}$ is a null vector, so its spinor equivalent is of the form $r\xi^A\bar{\xi}^{B'}$. Also the spinor equivalent of $\mathbf{OQ}$ has the form $th^{AB'}$. The equation $\mathbf{OP} = \mathbf{OQ} + \mathbf{QP}$ (see figure 1) gives

$$x^{AB'} = th^{AB'} + r\xi^A\bar{\xi}^{B'}. \tag{5·3}$$

In order to fix the choice of ξ^A, and hence also to fix r, let us choose a constant spinor η^A satisfying

$$h^{AB'}\eta_A\bar{\eta}_B = 1 \tag{5·4}$$

and so normalize ξ^A by

$$\xi_A \eta^A = 1. \tag{5·5}$$

There will then be just one ξ^A corresponding to each null direction (with the exception of that corresponding to η^A). Define the complex number ζ by

$$h^{AB'}\xi_A\bar{\eta}_{B'} = \zeta. \tag{5·6}$$

Then multiply (5·2) by $\xi^A\bar{\xi}^{B'}\eta^C\bar{\eta}^{D'}$ to give

$$h^{AB'}\xi_A\bar{\xi}_B = 1+\zeta\bar{\zeta}. \tag{5·7}$$

by (5·4), (5·5) and (5·6).

Now apply $\nabla_{CD'}$ to (5·3) to give

$$\delta^A_C\delta^{B'}_{D'} = h^{AB'}\nabla_{CD'}t+\xi^A\bar{\xi}^{B'}\nabla_{CD'}r+r\bar{\xi}^{B'}\nabla_{CD'}\xi^A+r\xi^A\nabla_{CD'}\bar{\xi}^{B'} \tag{5·8}$$

by (5·1) ($h^{AB'}$ is constant). Multiply (5·8) by $\xi_A\bar{\xi}_{B'}$ and use (5·7) to obtain

$$\nabla_{CD'}t = (1+\zeta\bar{\zeta})^{-1}\xi_C\bar{\xi}_{D'}. \tag{5·9}$$

Substitute (5·9) back in (5·8) and multiply by $\xi_A\bar{\eta}_{B'}$:

$$r\xi_A\nabla_{CD'}\xi^A = \xi_C\bar{\lambda}_{D'}, \tag{5·10}$$

where

$$\lambda_A = \bar{\zeta}(1+\zeta\bar{\zeta})^{-1}\xi_A-\eta_A \tag{5·11}$$

(see (5·6)). But η_A and $\eta_A\xi^A$ are constant (see (5·5)), so $\eta_A\nabla_{CD'}\xi^A = 0$. Combining this with (5·10) gives

$$\nabla_{CD'}\xi^A = r^{-1}\xi_C\bar{\lambda}_{D'}\eta^A \tag{5·12}$$

whence from (5·6), (5·4),

$$\nabla_{CD'}\zeta = r^{-1}\xi_C\bar{\lambda}_{D'}. \tag{5·13}$$

Multiply (5·8) by $\eta_A\bar{\eta}_{B'}$ and use (5·4), (5·9):

$$\nabla_{CD'}r = \eta_C\bar{\eta}_{D'}-(1+\zeta\bar{\zeta})^{-1}\xi_C\bar{\xi}_{D'}. \tag{5·14}$$

Observe that the variables $t, r, \zeta, \bar{\zeta}$ may be used as co-ordinates for space-time†, ζ being a complex angular parameter (related to spherical polars by $\zeta = e^{i\phi}\cot\frac{1}{2}\theta$), $\sqrt{2}t$ a retarded time parameter and $r(1+\zeta\bar{\zeta})/\sqrt{2}$ a radial distance parameter. I shall be concerned with the asymptotic behaviour of functions of $t, r, \zeta, \bar{\zeta}$ as r approaches infinity along some line t = const., ζ = const. In order to be able to differentiate order of magnitude symbols, the following notation will be used for a function $f(t, r, \zeta, \bar{\zeta})$ of '*asymptotically smooth*' order r^{-k}:

$$f = O^*(r^{-k})$$

along some line g, given by t = const., ζ = const. This will be taken to mean that not only

$$f = O(r^{-k}) \tag{5·15}$$

along g for large r, but also,

$$\frac{\partial f}{\partial r} = O(r^{-k-1}); \quad \frac{\partial f}{\partial\zeta}, \frac{\partial f}{\partial\bar{\zeta}}, \frac{\partial f}{\partial t} = O(r^{-k}) \tag{5·16}$$

† The quantities introduced here may be related to a standard Minkowski co-ordinate system (T, X, Y, Z) by:

$$\begin{pmatrix} t+r & r\bar{\zeta} \\ r\zeta & t+\zeta\bar{\zeta}r \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} T-Z & X-iY \\ X+iY & T+Z \end{pmatrix} = x^{AB'}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \eta^A, \begin{pmatrix} 1 \\ \zeta \end{pmatrix} = \xi^A.$$

along g (for large r). Similarly,

$$\frac{\partial^2 f}{\partial r^2} = O(r^{-k-2}); \quad \frac{\partial^2 f}{\partial r\,\partial\zeta}, \ldots = O(r^{-k-1}); \frac{\partial^2 f}{\partial\zeta\,\partial t}, \ldots = O(r^{-k}) \tag{5·17}$$

and so on, for as many derivatives as is required for the particular calculation under consideration. Alternatively, we can simply insist on (5·15), (5·16), (5·17), and so on, for *arbitrarily many* derivatives.

The reason for considering such 'asymptotically smooth' functions (see Newman & Penrose 1962; Sachs 1962*b*) is that they enable the essential differential operations to be carried out without a more restrictive assumption, such as of a power series in r^{-1}, having to be made (compare Bondi *et al.* 1962; Sachs 1962*a*). If a function of position is to be expandable in powers of r^{-1} for *every* choice of origin O, then it must necessarily be analytic in *all* the variables r, t, $\zeta \pm \bar{\zeta}$. This would exclude the possibility of such 'time bomb' functions as considered by Bondi *et al.* On the other hand, it is not hard to see that 'asymptotic smoothness' in all directions, is independent of the position of the origin. Note also that $r^{-k-\epsilon}\ln r(\epsilon > 0)$, $r^{-k}(\ln r)^{-1}$, $r^{-k}\sin\ln r$, $e^{-r}\sin r$, etc., are all $O^*(r^{-k})$, although $r^{-k}\sin r$ is only $O(r^{-k})$. None of them is expandable in powers of r^{-1}. It is important to observe, moreover, the *physical* implication of a condition such as $\chi = O^*(r^{-1})$ on a solution of the wave equation $\Box\chi = 0$. This would imply, for example (unlike $\chi = O(r^{-1})$), that any incoming waves would have to fall off suitably either in amplitude or in frequency. No incoming periodic solution could satisfy $\chi = O^*(r^{-1})$.

Suppose $f_{\cdots}$ is some spinor function of position, where

$$f_{\cdots} = O^*(r^{-k})$$

along g. Then since the spinor gradient of $f_{\cdots}$ has the form

$$\nabla_{CD'} f_{\cdots} = \frac{\partial f_{\cdots}}{\partial r}\nabla_{CD'} r + \frac{\partial f_{\cdots}}{\partial\zeta}\nabla_{CD'}\zeta + \frac{\partial f_{\cdots}}{\partial\bar{\zeta}}\nabla_{CD'}\bar{\zeta} + \frac{\partial f_{\cdots}}{\partial t}\nabla_{CD'} t, \tag{5·18}$$

we see from (5·9), (5·13), (5·14) and (5·12) that

$$\nabla_{CD'} f_{\cdots} = O^*(r^{-k}) \tag{5·19}$$

along g and furthermore, that

$$\xi^C \nabla_{CD'} f_{\cdots} = O^*(r^{-k-1}) \tag{5·20}$$

along g.

6. Asymptotic behaviour in flat space-time

The potentials introduced in §4 and the results and notation of §5 can now be combined to give a short proof, for flat space-time, of the *peeling-off* property of the p.n.d. mentioned in the introduction. An inductive argument obtains the peeling-off property for a spin s zero rest-mass field based on the assumption that the field is locally expressible (by means of (4·1)) in terms of a spin $(s-\frac{1}{2})$ field which itself satisfies the appropriate peeling-off condition asymptotically smoothly. Ultimately the assumption comes down to the fact that the Hertz-type potential χ should be $O^*(r^{-1})$.

Let $\phi_{AB\ldots K}$ be a symmetric spinor with $2s$ indices satisfying the spin s zero rest-mass equation (2·5). The required peeling-off property will now be stated more precisely as follows:

$$\phi_{AB\ldots FG\ldots K}\xi^G\ldots\xi^K = O(r^{-k-1}) \tag{6·1}$$

along all lines g given by $t = \text{const.}$, $\zeta = \text{const.}$, where there are k ξ's on the left with $k = 0, 1, \ldots, 2s$. For (6·1) implies that to order r^{-k} the left-hand side vanishes, so by (3·4), at least $2s-k+1$ p.n.d. coincide to that order, in the (radial) null direction corresponding to ξ^A. Note, in passing, the equivalent way of expressing (6·1):

$$\phi_{A\ldots K} = \sum_{j=0}^{2s} \frac{\alpha_j}{r^{j+1}}\xi_{(A}\ldots\xi_F\eta_G\ldots\eta_{K)},$$

where there are $2s-j$ ξ's and j η's and where each α_j is bounded as $r\to\infty$ (given t, ζ). Also if the assumption of a power series in r^{-1} is made, other equivalent forms are possible.

To prove (6·1) for a particular line $t = t_0$, $\zeta = \zeta_0$, let us, in the neighbourhood of this line, put

$$\phi_{AB\ldots K} = \mu^{M'}\nabla_{AM'}\psi_{B\ldots K} \tag{6·2}$$

as in (4·1), where $\mu^{M'}$ is constant and where $\psi_{B\ldots K}$ is symmetric, satisfies the spin $(s-\frac{1}{2})$ equation and, as an inductive hypothesis, satisfies

$$\psi_{B\ldots GH\ldots K}\xi^H\ldots\xi^K = O^*(r^{-k}) \tag{6·3}$$

along g there being $k-1$ ξ's with $k = 1, 2, \ldots, 2s$. Operating on (6·3) with $\xi^A\mu^{M'}\nabla_{AM'}$ leads at once to (6·1) with $k \neq 0$, by (5·20), since from (5·12) we have $\xi^A\nabla_{AM'}\xi^H = 0$. It remains to prove (6·1) in the case $k = 0$. This can be simply achieved by operating on (6·3), in the case $k = 1$, with $\mu^{M'}\nabla_{AM'}$ and using (5·19). Finally, the result rests on the condition that the Hertz-type potential χ should satisfy

$$\chi = O^*(r^{-1}), \tag{6·4}$$

along g. This seems 'reasonable' physically, if the incoming waves tail off suitably, although there is no *a priori* reason why such a physically reasonable field should be obtainable from such a physically reasonable Hertz-type potential. However, an assumption *has* to be made at some stage, since there clearly are 'non-physical' fields (e.g. a constant field, or a plane or cylindrical wave) which do not satisfy (6·1). In fact χ need not (and usually will not) be 'physically reasonable', in all respects other than (6·4), if considered from a *global* point of view (see § 7).

It is of some interest to note the strengthening of (6·1) that can be obtained in particular cases. For example, if in some directions χ falls off more rapidly than r^{-1}, say as $O^*(r^{-j-1})$, then the above proof carries through to give the stronger result that in these directions (6·1) holds with $k+j$ in place of k on the left-hand side. More generally, if any of the intermediate potentials satisfies such a strengthened version of (6·1) (for any particular g, k say), then the corresponding property will carry through to the final field $\phi_{A\ldots K}$. Another result which may be regarded as a kind of limiting case of this, but which is at any rate immediate from the above proof, is the following: *if, everywhere, $\psi_{B\ldots K}$ has the radial† null direction*

† If $k \geqslant 2$, this result can be slightly generalized, the restriction that the relevant null direction be radial being unnecessary.

(corresponding to ξ_A) *as a k-fold p.n.d., then so has* $\phi_{AB\ldots K}$. A further case of interest is given if, instead, $\psi_{B\ldots K}$ is *static* (i.e. independent of t). For then (by (5·18), etc.) $\mu^{M'}\nabla_{AM'}$ (and not just $\xi^A\mu^{M'}\nabla_{AM'}$) will reduce the order of $\psi_{B\ldots M}$ by unity and so $\phi_{AB\ldots M} = O(r^{-2})$. If some earlier potential is static the order of $\phi_{A\ldots M}$ would be reduced correspondingly further.

Most of the above points will be illustrated in the special examples to be considered shortly. However, there is one asymptotic feature which holds *generally* for fields represented by (6·2) and this concerns the null direction corresponding to $\mu^{M'}$. If, in fact, we choose $\xi^A = \bar{\mu}^M$, then $\mu^{M'}\nabla_{AM'}$ reduces the order of $\psi_{B\ldots K}$ by unity (see (5·20)) so we see that $\phi_{A\ldots K} = O(r^{-2})$ in this direction in general. Repeating this argument we see that if $\phi_{A\ldots K}$ is represented by (4·3) (and (6·4) holds), then in each of the $2s$ directions given by $\mu^{M'}, \nu^{M'}, \ldots, \omega^{M'}$ the field falls off as r^{-2} rather than r^{-1}. That is to say the *radiation field* (i.e. the 'null part' of the field) vanishes in these directions. If there are multiplicities among $\mu^{M'}, \nu^{M'}, \ldots, \omega^{M'}$, then the field falls off as r^{-1-j} where j is the relevant multiplicity.

Such behaviour is in fact not so special as it might seem. For there are topological reasons that, in general, there is a net algebraic count of $2s$ null directions along each cone $t = \text{const.}$ in which the radiation field must vanish, that is, in which the field must fall of as r^{-2}. The thing which *is*, however, special about the above behaviour is that the null directions given by $\mu^{M'}, \nu^{M'}, \ldots, \omega^{M'}$ must be *constant* in time, whereas the directions in which the radiation field must vanish may vary in time. This gives an additional argument for regarding (6·2) as being only a *local* representation of the field, which can be extended to infinity, in general, only in null directions distinct from that corresponding to $\mu^{M'}$. In order to cover the whole of space-time several different potentials and different choices of $\mu^{M'}$ would generally have to be made.

It may be mentioned in passing that there is a generalization of (6·2) which avoids this difficulty:

$$\phi_{AB\ldots K} = (\mu^{M'} + x^{LM'}\gamma_L)\nabla_{AM'}\psi_{B\ldots K} + 2s\gamma_{(A}\psi_{B\ldots K)}$$

where $\mu^{M'}$ and γ_L are both constant and where $x^{LM'}$ is the co-ordinate spinor. This more general expression has the added advantage that it transforms into another of the same form under conformal transformations of the Minkowski space-time. This has importance in relation to part II. A full discussion of these matters would take us too far afield for our present purposes, however, and will be given elsewhere.

7. Some examples in flat space-time

Using the notation of § 5, put $$R = r(1+\zeta\bar{\zeta}). \tag{7·1}$$

Then since $R/\sqrt{2}$ measures radial distance, the spin zero monopole can be written

$$\chi = aR^{-1}, \tag{7·2}$$

where a is a constant. Applying (4·3) to this gives the general spin s 2^s-pole:

$$\phi_{AB\ldots K} = a\mu^{M'}\nu^{N'}\ldots\omega^{W'}\nabla_{AM'}\nabla_{BN'}\ldots\nabla_{KW'}R^{-1} \tag{7·3}$$

$\mu^{M'}, \ldots, \omega^{W'}$ being constant (see also Sachs & Bergmann 1958).

As a particular case, the electromagnetic dipole is given by

$$\phi_{AB} = a\mu^{M'}\nu^{N'}\nabla_{AM'}\nabla_{BN'}R^{-1}. \tag{7·4}$$

If $\mu^{M'}$ and $\nu^{N'}$ happen to correspond to diametrically opposite null directions (relative to the time axis h), i.e. $h_{AB'}\overline{\mu}^{A}\nu^{B'} = 0$, then (7·4) gives a pure electric dipole, or a pure magnetic dipole with the same axis, or a mixture of these two, depending on the reality properties of the number $a\mu_{M'}\nu^{M'}$. At the other extreme is the case when $\mu^{M'}$ and $\nu^{N'}$ are proportional. This gives an equal mixture of magnetic and electric dipoles with perpendicular axes. In all there will be three linearly independent expressions (7·4) if complex coefficients are allowed (six with real coefficients).

For spin s there will be $2s+1$ independent expressions (7·3) (or $4s+2$ if real coefficients are used). (Compare Sachs & Bergmann 1958.) Note that s need not be integral in (7·3). Thus, for $s = \frac{1}{2}$ we have an interesting† and perfectly respectable static solution of Weyl's neutrino equation with a time-like source—the '$\sqrt{2}$-pole'.

If (7·2) is modified to incorporate outgoing radiation:

$$\chi = f(t)\,R^{-1}, \tag{7·5}$$

then the application of (4·3) gives us a spin s radiating 2^s-pole. This is not the most general outgoing radiating 2^s-pole, however, for reasons mentioned at the end of §6. To obtain the most general form, either linear combinations of these expressions may be taken (see (4·9) *et seq.*) or other expressions must be used. (If (4·3) were to be always retained globally for a *single* choice of $\mu^{N'}, \ldots, \omega^{N'}$ and χ, then (6·4) would sometimes have to fail in certain directions for this χ.) To obtain higher multipoles and radiating higher multipoles, (7·5) must be replaced by its derivative to required order. Thus, for example,

$$\chi = \nabla_{LM'}\{A^{LM'}(t)\,R^{-1}\}$$

gives, via (4·3), an example of a spin s radiating 2^{s+1}-pole, $A^{LM'}(t)$ being a function of t only. Again, for the general form linear combinations may be taken. Higher multipoles are similar. In all these cases the p.n.d. must satisfy the peeling-off property in future null directions, since $\chi = O^*(r^{-1})$

The Hertz-type potentials for lower order multipoles are not quite so easily obtained. Consider first the case ($s = 1$) of a Coulomb field, for which put

$$\chi = -\tfrac{1}{2}at\zeta^{-1}r^{-1} \tag{7·6}$$

(a being constant). Then $\Box\chi = 0$, $\chi = O^*(r^{-1})$ and

$$\psi_A = \overline{\eta}^{M'}\nabla_{AM'}\chi = -\tfrac{1}{2}aR^{-1}\zeta^{-1}\xi_A, \tag{7·7}$$

$$\phi_{AB} = \overline{\eta}^{M'}\nabla_{AM'}\psi_B = aR^{-2}\xi_{(A}\lambda_{B)}, \tag{7·8}$$

as is easily verified by using (5·9), (5·14), (7·1), (5·13) and (5·12). To see that (7·8) represents a Coulomb-type field (electric, magnetic or a mixture of two, depending on the reality properties of a) we observe that it is static and has the right dependence on R and that the p.n.d.—corresponding respectively to ξ_A and λ_B—are the outward radial null direction and the opposite inward null direction. (That

† The p.n.d. are the tangents to set a of right-circular helices which pass infinitely many times through the singular time-like line.

these null directions are, in fact, opposite relative to the time-axis h is implied by $h^{AB'}\xi_A\bar{\lambda}_{B'} = 0$, which follows from (5·11), (5·6), (5·7), and (5·5). Also we have the normalization $\lambda_A\xi^A = 1$. The spherical symmetry of the expression (7·8) is implicit in all this.

Note that the potential χ is singular in the direction $\zeta = 0$ and also in the direction $\zeta = \infty$ (since $\zeta\bar{\zeta}r$ is finite at $\zeta = \infty$, by (7·1) so ζr tends to zero.). The potential ψ_A is singular along $\zeta = 0$, but it turns out to be regular along $\zeta = \infty$. Note also the fact that χ depends on t. This is essential if we require, as here, $\chi = O^*(r^{-1})$. For, with a static χ with this asymptotic dependence, we should have $\psi_A = O^*(r^{-2})$ and static, whence ϕ_{AB} would have to be $O(r^{-3})$. (Static Hertz potentials for (7·8) are, however, possible with $\chi \neq O^*(r^{-1})$, for example $\chi = \frac{1}{2}a\zeta^{-1}$.)

For the spin 2 monopole (linearized Schwarzschild solution) we may put

$$\chi = \tfrac{1}{12}at^2\,(2t - 3R)\,\zeta^{-2}r^{-2}. \tag{7·9}$$

This is again $O^*(r^{-1})$ with $\Box\chi = 0$ and is singular along $\zeta = 0$ and $\zeta = \infty$. Differentiating, we get

$$\psi_{AB} = \bar{\eta}^{M'}\bar{\eta}^{N'}\nabla_{AM'}\nabla_{BN'}\chi = \tfrac{1}{12}aR^{-1}\zeta^{-2}\xi_A\xi_B, \tag{7·10}$$

which is singular only along $\zeta = 0$ and thence†

$$\phi_{ABCD} = \bar{\eta}^{M'}\bar{\eta}^{N'}\nabla_{AM'}\nabla_{BN'}\psi_{CD} = aR^{-3}\xi_{(A}\xi_B\lambda_C\lambda_{D)}, \tag{7·11}$$

which is static and has the correct dependence on R, and has spherical symmetry, for the same reasons as before. The p.n.d. coincide in pairs one pair pointing radially outwards and the other pair radially inwards as is the case with Schwarzschild's solution of the full gravitational equations (Penrose 1960). The expressions for the higher spin monopoles appear to be essentially similar to those already given. Also in order to construct a spin s 2^{s-k}-pole for $0 < k < s$, we have only to use a spin k monopole solution as a potential. Thus, in particular, the general spin s 2^{s-1}-pole and 2^{s-2}-pole solutions, respectively, may be obtained from (7·6) and (7·9) by means of (4·3) where, in the first case two, and in the second case four of $\mu^{M'}$, $\nu^{M'}, \ldots, \omega^{M'}$ are put equal to $\bar{\eta}^{M'}$.

Of course the choice of Hertz-type potential, etc., for all these solutions is very far from unique. For example, any $O^*(r^{-1})$ function of $t+r$ and ζr can be added to (7·6) or (7·9) without essential change. Also the choice of $\bar{\eta}^{M'}$ in (7·7), (7·8), (7·10) and (7·11) is clearly only a matter of convenience. Suppose, for example, that the $\bar{\eta}$'s of (7·8) and (7·11) are replaced by the constant spinor $\bar{\xi}^{Z'} - \bar{\zeta}\bar{\eta}^{Z'}$ which corresponds to the null direction opposite to that of η^Z. Then the resulting fields ϕ_{AB} and ϕ_{ABCD} are unchanged if (7·6) and (7·9) are respectively replaced by $\chi = \frac{1}{2}a\ln(1+tr^{-1})$ and $\chi = \frac{1}{12}a\{R\ln(1+tr^{-1}) - t(t+R)\,(t+r)^{-1}\}$, both of which are $O^*(r^{-1})$ and satisfy $\Box\chi = 0$. They are singular along $z = \infty$ and also along the hyperplane $t+r = 0$. If *all* occurrences of $\bar{\eta}^{Z'}$ in (7·7), (7·8), (7·10) (7·11) are replaced by $\bar{\xi}^{Z'} - \bar{\zeta}\bar{\eta}^{Z'}$, then, curiously enough, the complex conjugates of the original potentials (7·6) and (7·9) will work with only a change of sign in (7·8). The presence of singularities in the χ's

† These calculations are, perhaps, somewhat facilitated if carried out in terms of t, $\zeta r, R$, $\zeta^{-1}\xi_A, \zeta\lambda_A$. Note, also that terms of $t, R, \zeta, \bar{\zeta}$, we have

$$\Box \equiv R^{-1}\partial(R\partial/\partial t)/\partial R - R^{-2}\partial(R^2\partial/\partial R)\,\partial R - (1+\zeta\bar{\zeta})^2\,R^{-2}\,\partial^2/\partial\zeta\partial\bar{\zeta}$$

and also in ψ_A and ψ_{AB} is apparently *essential* in all these cases since otherwise the expression for mass or charge would have to vanish (Sachs & Bergmann 1958).

Observe that the passage from (7·7) to (7·8) or from (7·10) to (7·11) illustrates a result mentioned in § 6. There is a radial p.n.d. in each case whose multiplicity carries through from the potential to the resulting field. Also in the intermediate fields (7·7) and (7·10) belong to a class of null solutions of (2·5) which are of considerable interest in themselves. A more general example is given by

$$R^{-1}f(t,\zeta)\,\xi_A\xi_B\ldots\xi_K, \tag{7·12}$$

where f is an arbitrary function of t and ζ which is analytic in ζ. Such solutions (and others which are not centred on a time-like straight line) have been studied by Bateman (1955) and by Robinson & Trautman (1962). Unfortunately, if the relevant more general expression (7·12) is used in place of ψ_A or ψ_{AB} in (7·8), (7·11) then the extra singularities will remain the resulting fields.

Null free-wave solutions for each spin can, however, be constructed which, unlike (7·12), are everywhere non-singular and which satisfy the peeling-off property in all directions. (It is, in fact, then a consequence that these waves have finite total energy, angular momentum, etc.) Furthermore, similar solutions of *any* other specified algebraic type can also be exhibited. Put

$$\chi = \frac{a}{(\theta_\mu + x_\mu)\,(\theta^\mu + x^\mu)},$$

where a, is a constant scalar and where θ^μ is a constant complex vector with a *time-like imaginary part*. Evidently χ is nowhere singular and is $O^*(r^{-1})$ in all directions (see Synge 1956, p. 361). We have

$$\nabla_\nu\chi = -2\chi^2 a^{-1}(\theta_\nu + x_\nu), \tag{7·13}$$

so that

$$\Box\chi = 0.$$

If we put

$$\alpha_A = \tfrac{1}{2}\chi^{-2}a\mu^{M'}\nabla_{AM'}\chi = -\mu^{M'}\theta_{AM'} - \mu^{M'}x_{AM'} \tag{7·14}$$

(from 7·13), where $\mu^{M'}$ is constant, then

$$\nabla_{BL'}\alpha^A = \delta^A_B\mu_{L'}. \tag{7·15}$$

By (7·14), the spin $\frac{1}{2}$ field $\psi_A = \mu^{M'}\nabla_{AM'}\chi$ satisfies $\alpha^A\psi_A = 0$. Similarly, if we repeat this according to (4·2) we get a spin $2s$ field

$$\phi_{AB\ldots K} = \mu^{M'}\mu^{N'}\ldots\mu^{W'}\nabla_{AM'}\nabla_{BN'}\ldots\nabla_{KW'}\chi,$$

which is evidently non-singular (and non-zero) and which satisfies (6·1) in all directions. Since from (7·15) the operators $\alpha^A\mu^{M'}\nabla_{AM'}$ and $\mu^{N'}\nabla_{BN'}$ commute, it follows by (7·14) that

$$\alpha^A\phi_{AB\ldots K} = 0,$$

whence $\phi_{AB\ldots K}$ is *null* with $2s$-fold p.n.d.† corresponding to α^A.

If we use (4·3) in place of (4·2):

$$\phi_{AB\ldots K} = \mu^{M'}\nu^{N'}\ldots\omega^{W'}\nabla_{AM'}\nabla_{BN'}\ldots\nabla_{KW'}\chi,$$

† The p.n.d. are tangential to a non-singular shear-free twisting congruence of null straight lines of a type discovered by I. Robinson (unpublished). Similar twisting null fields with some of the properties of those given here were also suggested by Robinson.

where $\mu^{M'}, \nu^{N'}, \ldots, \omega^{W'}$ are constant, then,

$$\alpha^A\alpha^B\ldots\alpha^K\phi_{AB\ldots K} = 0 = \beta^A\beta^B\ldots\beta^K\phi_{AB\ldots K} = \ldots = \kappa^A\kappa^B\ldots\kappa^K\phi_{AB\ldots K},$$

where $$\beta^A = \tfrac{1}{2}\chi^{-2}a\nu^{M'}\nabla^A_{M'}\chi, \quad \ldots, \quad \kappa^A = \tfrac{1}{2}\chi^{-2}a\omega^{M'}\nabla^A_{M'}\chi$$

since with (7·15) $$\nabla_{BL'}\beta^A = \delta^A_B\nu_{L'}, \quad \ldots, \quad \nabla_{BL'}\kappa^A = \delta^A_B\omega_{L'}.$$

The p.n.d. of $\phi_{AB\ldots K}$ therefore correspond to $\alpha^A, \beta^A, \ldots, \kappa^A$, and using a limiting argument, we see that coincidences among the p.n.d. of $\phi_{AB\ldots K}$ arise precisely from the coincidences (up to proportionality) among $\mu^{M'}, \nu^{M'}, \ldots, \omega^{M'}$. Thus, the algebraic type (coincidence scheme for the p.n.d.) of $\phi_{AB\ldots K}$ is the same at *every* event, being the same as that for $\mu^{M'}, \ldots, \omega^{M'}$. It is perhaps instructive to note that the peeling-off property (to be understood in the sense of (6·1)) holds despite this!

II. THE CONFORMAL TECHNIQUE

8. Conformal structure of Minkowski infinity

The arguments involving potentials used in part I do not appear to generalize in any obvious way which would enable them to be applied in curved space-times. However, part II will be devoted to a description of a different technique which can be used in *both* flat and curved space-times. The resulting agreement with § 6 that will be obtained (Sachs's peeling-off property) would seem to be a strong endorsement of the validity of the few simple assumptions made.

The study of asymptotic behaviour is the study of behaviour in the 'neighbourhood of infinity'. Actual points *at* 'infinity' are, however, not generally thought of as having any consistently sensible existence. The main idea here is to formalize a concept of 'points at infinity' so that they can, in fact, be introduced in a rigorous and consistent way. These points at 'infinity' will form a kind of hypersurface boundary to space-time. 'Asymptotic behaviour' (as considered here) will then turn out to be essentially no different from behaviour at any finite hypersurface. The basic tool is provided by the concept of a *conformal transformation* of space-time. Such transformations preserve *ratios* of local infinitesimal distances, or equivalently, the positions of the light cones. As a consequence, it will turn out that the essential features of zero rest-mass fields are preserved, and their asymptotic behaviour may be studied by using *local* differential techniques in the way indicated above.

One starting-point in the study of the conformal structure of infinity for *Minkowski* space-time is provided by the *inversion*:

$$\tilde{x}^\mu = -x^\mu(x_\nu x^\nu)^{-1}, \quad x^\mu = -\tilde{x}^\mu(\tilde{x}_\nu\tilde{x}^\nu)^{-1}; \tag{8·1}$$

$$\tilde{g}_{\mu\nu} = g_{\mu\nu} = \mathrm{diag}\{+1, -1, -1, -1\}.$$

This is conformal since $$g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu = \Omega^2\tilde{g}_{\mu\nu}\mathrm{d}\tilde{x}^\mu\mathrm{d}\tilde{x}^\nu,$$

where $$\Omega = x_\mu x^\mu = (\tilde{x}_\mu\tilde{x}^\mu)^{-1}.$$

The null cone of the x^μ origin is mapped to $\tilde{x}^\mu$ infinity and x^μ infinity is mapped into the null cone of the $\tilde{x}^\mu$ origin. (The minus sign in (8·1) is to preserve time sense.) We may regard (8·1) either as giving a conformal *point* transformation in Minkowski space-time which maps infinity into a finite null cone, or else, as a relation between

two *co-ordinate systems*, each of which covers part of a certain highly symmetrical (compact) conformal manifold $\mathscr{R}$ (Rudberg 1957; Penrose 1962), the 'compactification' of Minkowski space-time†. These two co-ordinate systems do not quite cover the whole of $\mathscr{R}$ there being a additional sphere of points lying at infinity on the null cones of both origins. But these remaining points are easily covered by choosing a third co-ordinate system which is related to $x^\mu - a^\mu$ in the same way as $\tilde{x}^\mu$ is related to x^μ, where a^μ is a fixed time-like vector. The resulting manifold $\mathscr{R}$ has many interesting properties, but they will not be discussed here, the essential relevant feature being that $\mathscr{R}$ is obtained from Minkowski space-time by the addition of a (closed) *null cone* $\mathscr{I}$ at infinity. From the point of view of conformal structure, $\mathscr{I}$ is just one null cone on an equal footing with any other null cone in $\mathscr{R}$ (all of which are closed).

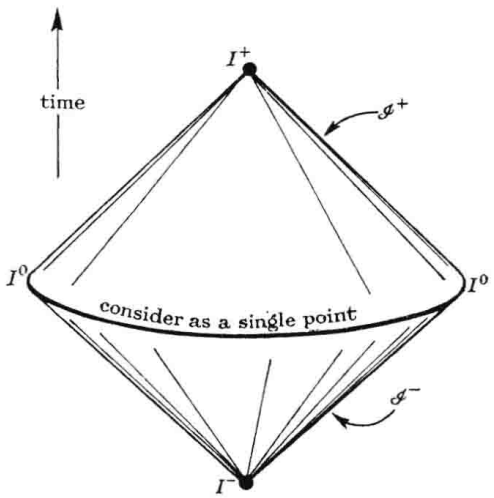


FIGURE 2. Minkowski space-time can be mapped onto the interior of a region bounded by two cones joined base to base. The 'equator' at which the cones are joined should be considered as a single point. To obtain the manifold $\mathscr{R}$, opposite generators must be identified with future sense preserved.

There is one property of $\mathscr{R}$, however, which seems undesirable when these ideas are applied to interacting fields, or to curved space-times. This is the fact that the 'future infinity' turns out to have been identified with the past infinity' in the definition of $\mathscr{R}$. To avoid this feature it will be desirable effectively to 'cut' this manifold along the hypersurface $\mathscr{I}$ and to consider instead the resulting manifold with boundary. This boundary consists essentially of two copies of $\mathscr{I}$, one in the 'future' which will be called $\mathscr{I}^+$ and one in the 'past' to be called $\mathscr{I}^-$ (see figure 2). The four-dimensional region between these two bounding hypersurfaces corresponds to the original Minkowski space-time. The hypersurface $\mathscr{I}^-$ is a 'closed up' null cone with a past vertex I^- and also a future vertex I^0. This point I^0 is also the past vertex of the null cone $\mathscr{I}^+$ and $\mathscr{I}^+$ also has a future vertex I^+. For preciseness, I shall exclude the three points I^-, I^0, I^+ from the definitions of the sets $\mathscr{I}^-$, $\mathscr{I}^+$.

† $\mathscr{R}$ is effectively a real projective quadric fourfold of signature $(+,+,-,-,-,-)$. See also Kuiper (1949).

In order to get a clearer picture of the situation, let us consider co-ordinates t, $\zeta, \bar{\zeta}, w$ in Minkowski space-time, where

$$w = t + r(1+\zeta\bar{\zeta})$$

with t, r, ζ defined as in § 5. Then $\sqrt{2}\,w$ and $\sqrt{2}t$ are advanced and retarded time parameters respectively and ζ is a complex angular parameter (related to the more usual spherical polar angles by $\zeta = e^{i\phi}\cot\frac{1}{2}\theta$). Thus the Minkowski metric $d\tilde{s}$ is given by

$$d\tilde{s}^2 = 2dw\,dt - 2(w-t)^2(1+\zeta\bar{\zeta})^{-2}\,d\zeta d\bar{\zeta},$$

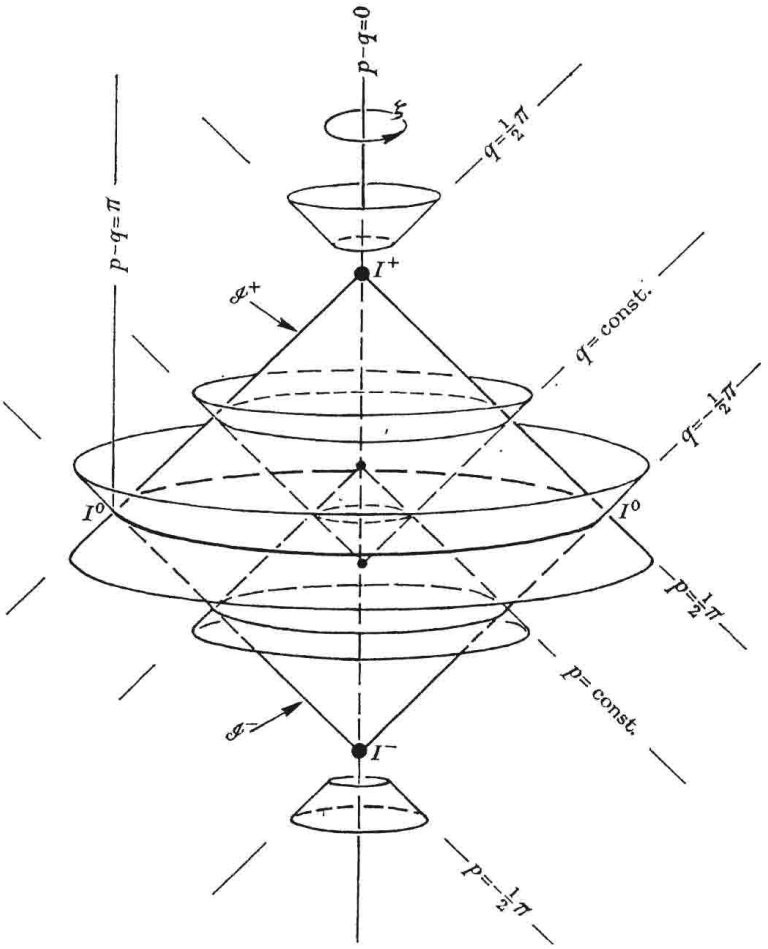


FIGURE 3. The relation of the co-ordinates for (8·3) to the picture of figure 2.

$w - t \geqslant 0$. Choose a new ('unphysical') metric ds with

$$ds^2 = \Omega^2 d\tilde{s}^2,$$

conformal to this where for convenience

$$\Omega = (1+w^2)^{-\frac{1}{2}}(1+t^2)^{-\frac{1}{2}}. \tag{8·2}$$

Then

$$ds^2 = 2\,dp\,dq - 2\sin^2(p-q)\,(1+\zeta\bar{\zeta})^{-2}\,d\zeta d\bar{\zeta}, \tag{8·3}$$

where

$$w = \tan p, \quad t = \tan q, \tag{8·4}$$

and the range of variation of p and q is from $-\frac{1}{2}\pi$ to $\frac{1}{2}\pi$ with $p-q \geqslant 0$. The hypersurface $\mathscr{I}^-$ is then given by $q = -\frac{1}{2}\pi$, $-\frac{1}{2}\pi < p < \frac{1}{2}\pi$ and $\mathscr{I}^+$ is given by $p = \frac{1}{2}\pi$, $-\frac{1}{2}\pi < q < \frac{1}{2}\pi$. The point I^- is given by $p = q = -\frac{1}{2}\pi$, I^0 by $p = -q = \frac{1}{2}\pi$ and I^+ by $p = q = \frac{1}{2}\pi$ (see figure 3).

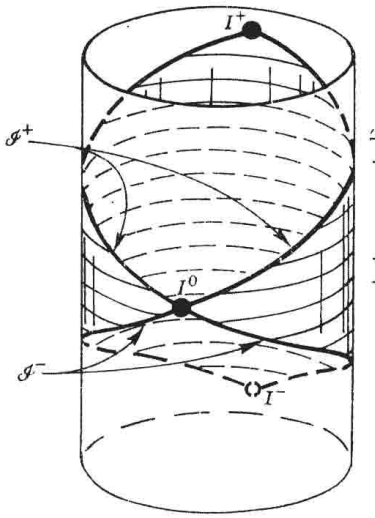


FIGURE 4

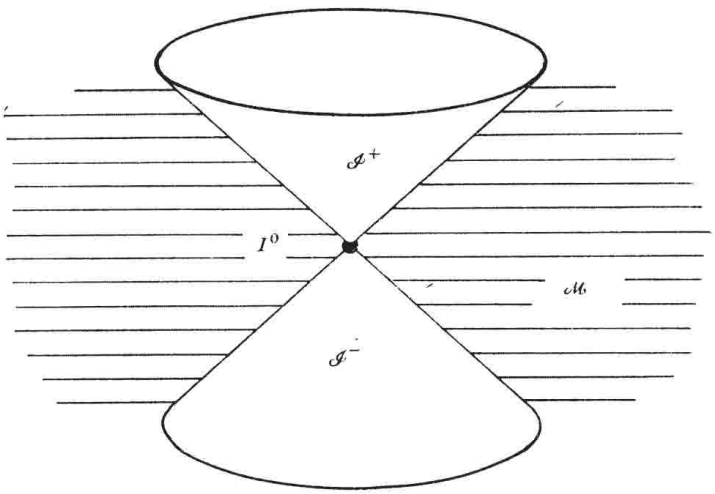


FIGURE 5

FIGURE 4. The cylinder represents the Einstein universe. The shaded region represents a part conformal to Minkowski space-time. (This time *two* space dimensions are suppressed in the diagram.)

FIGURE 5. The neighbourhood of I^0. Physical space-time is represented by the shaded region.

Observe that the metric (8·3) can be extended beyond the range of definition given above, provided merely $0 \leqslant p-q \leqslant \pi$ (the complex parameter ζ being unrestricted and with $\zeta = \infty$ allowed). Then (8·3) in fact represents the well-known static universe of Einstein: a spherical 3-space which is constant in time. These 3-space sections are given by $p+q = $ const. and the time-like lines orthogonal to these sections, by $p-q = $ const., $\zeta = $ const. This conformally flat model is everywhere non-singular. In particular if $p-q = 0$ or $p-q = \pi$ we get merely co-ordinate singularities representing two time-like lines of points diametrically opposite for the spherical 3-space. We thus see that the future null-cone $\mathscr{I}^-$ of the point I^- is refocused to a *point* (see figure 4), namely I^0, which is spatially diametrically opposite to it. Similarly, the future null cone $\mathscr{I}^+$ of I^0 is again refocused at I^+. We may, in fact, regard $\mathscr{I}^-$ and $\mathscr{I}^+$ as two parts of the *same* null cone, namely the past and future cone of I^0. In the neighbourhood of I^0 (see figure 5), Minkowski space-time corresponds to the *exterior* region of the cone.

The finite four-dimensional region lying between $\mathscr{I}^-$ and $\mathscr{I}^+$ is conformally identical with the original (infinite) Minkowski space-time; the finite boundary (I^-, $\mathscr{I}^-$, I^0, $\mathscr{I}^+$, I^+) of this region represents the 'infinity' of Minkowski space-time and is given by $\Omega = 0$. The points I^-, I^0 and I^+ can be interpreted respectively as the past, spatial and future infinities; the hypersurfaces $\mathscr{I}^-$, $\mathscr{I}^+$ represent past and future *null* infinities respectively. This may be verified by examining the behaviour at

infinity of straight lines in the original Minkowski space-time. If the line is time-like, then (after the conformal transformation) it meets the boundary in I^- and I^+ only; if null it meets it in one point of $\mathscr{I}^-$ and a corresponding point of $\mathscr{I}^+$; if the line is space-like, we get simply a closed curve through I^0. This behaviour will serve to indicate what should be expected also in curved asymptotically flat space-times.

9. Infinity for curved space-times

In order to decide on the appropriate asymptotic conditions to impose on a curved space-time, let us first examine Schwarzschild's solution. Co-ordinates t, ζ, R analogous to those of §§ 5 and 7 will be used and the metric (for mass $2^{-\frac{3}{2}}M$) takes the form

$$d\tilde{s}^2 = 2\,dt^2(1-MR^{-1}) + 2\,dR\,dt - 2R^2(1+\zeta\bar{\zeta})^{-2}\,d\zeta\,d\bar{\zeta}. \tag{9·1}$$

Thus, introducing $ds = \Omega\,d\tilde{s}$ with $\Omega = R^{-1}$and letting $l = R^{-1}$, we get,

$$ds^2 = 2\,dt^2(l^2 - Ml^3) - 2\,dl\,dt - 2(1+\zeta\bar{\zeta})^{-2}\,d\zeta\,d\bar{\zeta}. \tag{9·2}$$

This is manifestly regular (C^∞) in the neighbourhood of, and more particularly *at*, the (null) hypersurface $\mathscr{I}^+$, given by $l = 0$, t finite. Similarly, by re-expressing (9·1) (9·2) in terms of an advanced time parameter $w(= t + R + M\ln(R-M))$ in place of the retarded parameter t, we get regularity at $\mathscr{I}^-$, given by $l = 0$, w finite. The hypersurfaces $\mathscr{I}^+$, $\mathscr{I}^-$ (both given by $\Omega = 0$) are seen to be topologically three-dimensional open cylinders ($S^2 \times E^1$). This is precisely the same situation as with Minkowski space-time considered earlier (since the points I^-, I^0, I^+ are excluded from the definitions of $\mathscr{I}^-$, $\mathscr{I}^+$). With the conformal factor ($\Omega = R^{-1}$) given above, these cylinders extend to infinity at both ends. In order to make the *ends* of these cylinders finite as well, the conformal factor must be modified suitably, e.g. $\Omega = R^{-1}\{1 - \ln(e^{-t^2} + e^{-w^2})\}^{-\frac{1}{2}}$. This gives us three points I^-, I^0, I^+, just as with Minkowski space-time, where I^-, I^0, and I^0, I^+ respectively bound $\mathscr{I}^-$ and $\mathscr{I}^+$ to the past and future. However, there *is* an important new feature arising here, namely that the three points I^-, I^0, I^+ are all necessarily *singular* in the Schwarzschild case. This may be seen by examining the behaviour of the eigenvalues of the Weyl tensor for the transformed space-time (see § 10 and Pirani 1959). These scalars are proportional to $R^{-3}\Omega^{-2}M$ and in the neighbourhood of each of I^-, I^0, I^+ they are unbounded for any plausible choice of Ω ($\approx R^{-2}$ along $t = -w$) which makes the relevant point finite.

The metric (9·1) is a special case of

$$d\tilde{s}^2 = R^{-2}A\,dR^2 - 2B_i\,dx^i\,dR - R^2C_{ij}\,dx^i\,dx^j \tag{9·3}$$

($i, j = 1, 2, 3$) where if $R^{-1} = x^0$, then A, B_i, C_{ij} are to be regular (say C^3) functions of x^μ ($\mu = 0, 1, 2, 3$) on, and in the neighbourhood of, the hypersurface $\mathscr{I}$ defined by $x^0 = 0$. Assuming, also, that the relevant determinant in A, B_i, C_{ij} does not vanish, this clearly means that the metric $ds = \Omega\,d\tilde{s}$ with $\Omega = R^{-1}$ is regular (C^3) on $\mathscr{I}$. The metric (9·3) includes as particular cases many which have been suggested as representing gravitational radiation in asymptotically flat space-times (Robinson & Trautman 1962; Bondi 1960, 1962; Sachs 1962; and probably also Newman

& Unti 1962)†. The de Sitter metric also readily takes the form (9·3) (with $A = 1$, $B_i = 0, C_{ij} = \delta_{ij}$), where R is the exponential of the usual time parameter. Furthermore, some asymptotically de Sitter metrics (due to Demianski, private communication), which generalize those of Robinson-Trautman to include a cosmological term, also have the form (9·3). It would therefore seem that an assumption of regularity for $\mathscr{I}$ is not an unreasonable one to impose on a space-time if properties involving gravitational radiation are to be studied.

The choice of Ω for (9·3) has the important property that its gradient is non-vanishing ($\nabla_\mu \Omega \equiv \partial\Omega/\partial x^\mu = (1, 0, 0, 0)$ at $\mathscr{I}$) and it therefore determines the normal direction to $\mathscr{I}$ (since $\mathscr{I}$ is given by $\Omega = 0$). It will be useful here generally to supplement an assumption of regularity for $\mathscr{I}$ with this condition that the gradient of Ω be non-vanishing‡ there. However, it will be seen later that this condition can sometimes be deduced from the regularity assumption if certain additional physical assumptions are made.

To be more precise about the conditions to be satisfied, first, for $\tilde{\mathscr{M}}$ to be a *space-time*, we require on physical grounds that:

1. *$\tilde{\mathscr{M}}$ is a suitably differentiable, connected, orientable Riemannian 4-manifold without boundary, with metric* $\mathrm{d}\tilde{s}$ *of signature* $(+, -, -, -)$ *in which 'past' and 'future' null cones can be distinguished consistently and which contains no closed smooth timelike curves.*

Secondly, the space-time $\tilde{\mathscr{M}}$ will be called *asymptotically simple* if some manifold $\mathscr{M}$ with boundary $\mathscr{I}$ ($\subset \mathscr{M}$) and metric $\mathrm{d}s$ exists, whose interior $\mathscr{M} - \mathscr{I}$ is conformal to $\tilde{\mathscr{M}}$ with $\mathrm{d}s = \Omega\,\mathrm{d}\tilde{s}$ ($\Omega > 0$) and which satisfies the following properties:

2. *This manifold $\mathscr{M}$ is differentiable (say C^4) everywhere§, with regular (say C^3) metric* $\mathrm{d}s$. *(Regularity at points of $\mathscr{I}$ is to be interpreted as the possibility of regarding $\mathscr{M}$ as a part of a more extended four-dimensional manifold of which $\mathscr{I}$ is a set of regular interior points.)*

3. *Defining $\Omega = 0$ on $\mathscr{I}$, we have Ω regular (say C^3) throughout $\mathscr{M}$. Furthermore, $\nabla_\mu \Omega \neq 0$ on $\mathscr{I}$.*

4. *Every null geodesic in $\mathscr{M}$ contains, if maximally extended, two distinct points on $\mathscr{I}$.*

The reason for condition 4 is to ensure that $\mathscr{I}$ includes the *whole* of the 'null infinity' for $\tilde{\mathscr{M}}$. (It should be recalled that a null geodesic remains a null geodesic

† One would have to verify that the relevant C_{ij} are appropriately differentiable as functions of x^0, x^i. Thus, for example, we must have $\partial C_{ij}/\partial R = O(R^{-2})$, $2\partial C_{ij}/\partial R + R\partial^2 C_{ij}/\partial R^2 = O(R^{-3})$, etc.

‡ It has been pointed out by Norman (1964) that this condition necessarily fails for some cosmological models, notably for $\mathscr{I}^+$ in the Einstein-de Sitter universe. In the steady-state model, on the other hand, the condition $\Omega = 0$ fails for $\mathscr{I}^-$ (the space-time being incomplete in the past—'half' of a de Sitter space-time). In the exploding models, $\mathscr{I}^-$ can be treated as a non-singular hypersurface if we allow $\Omega = \infty$ there. Thus, the scope of the formalism can be somewhat extended by dropping the following condition 3 (see Penrose 1964), although it is necessary here.

§ Singular points such as the I^0, I^-, I^+ for Schwarzschild's solution should simply be omitted from the definition of $\mathscr{M}$. (However, the *mass* singularity must be 'smoothed out'.) Although it has a boundary, $\mathscr{M}$ will not usually be compact, i.e. there can be 'holes' in the boundary.

after conformal transformation, see also (13·1), (13·2), *et seq.*) Condition 4 is perhaps somewhat unnecessarily severe (and it may sometimes be difficult to verify in practice), since it rules out the possibility of light rays which do not escape to infinity. For example, there are the circular null orbits in Schwarzschild's solution at 3/2 Schwarzschild radii. Hence, in order to qualify to be 'asymptotically simple' as defined here, any space-time involving mass singularities must be replaced by one in which the masses are replaced by sufficiently extended regions of energy-momentum. Condition 4 does not, then, seem to imply any inappropriate loss in generality and it could be important if these ideas are to be applied in the construction of an 'S-matrix' type of theory incorporating gravitation. However, a weakened 'local' alternative version of condition 4 could perhaps be given if desired, although its precise statement might be considerably more involved.

The boundary set $\mathscr{I}$ of $\mathscr{M}$ must, of course, be three-dimensional. Furthermore, condition 3 implies that $\mathscr{I}$ is a (C^3) differentiable manifold, which is a *non-singular hypersurface* in $\mathscr{M}$. (Note, in particular, that in the Minkowski case, the vertices I^-, I^0, I^+ of $\mathscr{I}^-$, $\mathscr{I}^+$ must be omitted from the definition of $\mathscr{M}$, even though they would not be singular points of $\mathscr{M}$. According to (8·2), (8·4), $\Omega = \cos p \cos q$ whence $\nabla_\mu \Omega = 0$ at I^-, I^0, I^+.)

It is important to observe that the *conformal geometry* of $\mathscr{M}$ is completely determined by the physical space-time $\tilde{\mathscr{M}}$. If $\mathscr{M}$ and $\mathscr{M}'$ are related to $\tilde{\mathscr{M}}$ via $ds = \Omega d\tilde{s}$, $ds' = \Omega' d\tilde{s}$, respectively, then the *whole* of $\mathscr{M}$ is related by a non-singular conformal transformation $ds' = \Omega'' ds$ ($\Omega'' > 0$) to the *whole* of $\mathscr{M}'$. For we have $\Omega'' = \Omega' \Omega^{-1}$ in the interior region and $\Omega'' = \nabla_\mu \Omega' : \nabla_\mu \Omega$ on the boundary. (The vectors are proportional since they are both normal to the hypersurface $\mathscr{I}$.) Thus, $\mathscr{M}$ is physically meaningful if regarded as a *conformal* manifold. It is convenient for calculations however, to use a particular metric structure for $\mathscr{M}$ and to carry out covariant differentiations with respect to this metric. It will not be necessary that the particular metric chosen for a calculation be defined *globally*† on $\mathscr{M}$ if local conformal properties of $\mathscr{M}$ only are being investigated.

10. Conformal transformation formulae

Consider the conformal transformation $ds = \Omega d\tilde{s}$ ($\Omega > 0$), and suppose, for convenience, that the co-ordinates do not simultaneously undergo a transformation (contrast (8·1)). Then

$$\tilde{g}_{\mu\nu} = \Omega^{-2} g_{\mu\nu}, \quad \tilde{g}^{\mu\nu} = \Omega^2 g^{\mu\nu}. \tag{10·1}$$

Covariant derivative determined by the $\tilde{g}_{\mu\nu}$ metric (written $\tilde{\nabla}_\mu$) is given, in terms of the $g_{\mu\nu}$ covariant derivative (written ∇_μ), by

$$\tilde{\nabla}_\mu \theta = \nabla_\mu \theta, \tag{10·2}$$

$$\tilde{\nabla}_\mu k_\nu = \nabla_\mu k_\nu + 2\Omega^{-1} k_{(\mu} \nabla_{\nu)} \Omega - \Omega^{-1} g_{\mu\nu} k^\rho \nabla_\rho \Omega \tag{10·3}$$

† It might perhaps be argued that the definition of $\mathscr{M}$ should, more properly, have been given in terms of a 'patchwork' of metrics, with different Ω's for the various metrics. However, such a definition is easily seen to be equivalent to the one given here.

On the other hand, in agreement with Lichnerowicz' conditions, a set of weakened differentiability conditions might reasonably be given in terms of such a 'patchwork'.

(see, for example, Schouten 1954). Expression (10·3) generalizes in the usual way for covariant tensors with several indices; that is, $k_{\mu\ldots\rho}$ behaves as a product $m_\mu\ldots r_\rho$. Contravariant indices can be treated by lowering them with $g_{\mu\nu}$. Spinor indices will be dealt with shortly.

The sign of the curvature tensor will be chosen so that

$$2\nabla_{[\mu}\nabla_{\nu]}k_\rho = k_\sigma R^\sigma{}_{\rho\mu\nu}$$

and $\tilde{R}^\sigma{}_{\rho\mu\nu}$ is defined similarly using $\tilde{\nabla}_\mu$. Then, putting

$$P_{\mu\nu} = \tfrac{1}{2}R_{\mu\nu} - \tfrac{1}{12}Rg_{\mu\nu}, \tag{10·4}$$

where $R_{\mu\nu} = R^\sigma{}_{\mu\sigma\nu}$, $R = g^{\mu\nu}R_{\mu\nu}$, and defining $\tilde{P}_{\mu\nu}$ correspondingly, we get (compare Schouten 1954)

$$\tilde{P}_{\mu\nu} = P_{\mu\nu} - \Omega^{-1}\nabla_\mu\nabla_\nu\Omega + \tfrac{1}{2}\Omega^{-2}g_{\mu\nu}(\nabla_\rho\Omega)(\nabla^\rho\Omega). \tag{10·5}$$

The curvature information not contained in $P_{\mu\nu}$ (or, equivalently, $R_{\mu\nu}$) is given by Weyl's conformal tensor

$$C^{\mu\nu}{}_{\rho\sigma} = R^{\mu\nu}{}_{\rho\sigma} - 4P^{[\mu}{}_{[\rho}\delta^{\nu]}_{\sigma]},$$

which satisfies the Riemann tensor symmetries and $C^\mu{}_{\nu\mu\sigma} = 0$, and transforms according to

$$\tilde{C}_{\mu\nu\rho\sigma} = \Omega^{-2}C_{\mu\nu\rho\sigma}. \tag{10·6}$$

Note that when tensor indices are raised or lowered, all quantities with a tilde use $\tilde{g}^{\mu\nu}$, $\tilde{g}_{\mu\nu}$, while those without use $g^{\mu\nu}$, $g_{\mu\nu}$. Hence, for example $\tilde{C}^{\mu\nu}{}_{\rho\sigma} = \Omega^2 C^{\mu\nu}{}_{\rho\sigma}$ and $\tilde{C}^\mu{}_{\nu\rho\sigma} = C^\mu{}_{\nu\rho\sigma}$. Thus, the conformal *invariance* of $C^\mu{}_{\nu\rho\sigma}$ is contained in the fact that $C_{\mu\nu\rho\sigma}$ (or $C^{\mu\nu}{}_{\rho\sigma}$) is a suitable conformal *density*. Generally, a quantity $\xi_{\ldots}$ will be called a *conformal density of weight* $-m$ if

$$\tilde{\xi}_{\ldots} = \Omega^m\xi_{\ldots}.$$

In order to treat spinors, put

$$\tilde{\sigma}_\mu^{AB'} = \Omega^{-1}\sigma_\mu^{AB'}, \quad \tilde{\sigma}^\mu_{AB'} = \Omega\sigma^\mu_{AB'}. \tag{10·7}$$

This preserves (2·1) with (10·1). The ϵ's are defined to be unaffected by the conformal transformation: $\tilde{\epsilon}_{AB} = \epsilon_{AB}$, etc. The weight of a conformal density is, therefore, unaffected by the raising or lowering of *spinor* indices. In spinor terms the covariant derivative of a general spinor conformal density transforms (Buchdahl 1959) according to

$$\begin{aligned}\tilde{\nabla}_{XX'}\tilde{\xi}_{A\ldots LP'\ldots T'} = {}& \Omega^{m+1}\nabla_{XX'}\xi_{A\ldots LP'\ldots T'} + \Omega^m\{(m-\tfrac{1}{2}r)\,\xi_{A\ldots LP'\ldots T'}\nabla_{XX'}\Omega \\ & + \xi_{X\ldots LP'\ldots T'}\nabla_{AX'}\Omega + \ldots + \xi_{A\ldots XP'\ldots T'}\nabla_{LX'}\Omega + \xi_{A\ldots LX'\ldots T'}\nabla_{XP'}\Omega + \ldots \\ & + \xi_{A\ldots LP'\ldots Y'}\nabla_{XT'}\Omega\},\end{aligned} \tag{10·8}$$

where r is the total number of spinor indices of $\xi_{\ldots}$ and where

$$\tilde{\xi}_{A\ldots LP'\ldots T'} = \Omega^m\xi_{A\ldots LP'\ldots T'}.$$

To verify the correctness of (10·8), it is sufficient to observe that the sum and product rules for differentiation are preserved, that (10·8) reduces to (10·2) when $r = 0$, to (10·3) when $k_\mu = \sigma_\mu^{AP'}\xi_{AP'}$ with $m = 1$ ($r = 2$) and that $\tilde{\nabla}_{XY'}\epsilon_{AB} = 0$.

The spinor form of the curvature tensor decomposes as follows† (Penrose 1960)

$$R_{\mu\nu\rho\tau}\sigma^{\mu}_{AE'}\sigma^{\nu}_{BF'}\sigma^{\rho}_{CG'}\sigma^{\tau}_{DH'} = \Psi_{ABCD}\epsilon_{E'F'}\epsilon_{G'H'} + \epsilon_{AB}\epsilon_{CD}\bar{\Psi}_{E'F'G'H'}$$
$$+ 2\Lambda\{\epsilon_{AC}\epsilon_{BD}\epsilon_{E'F'}\epsilon_{G'H'} + \epsilon_{AB}\epsilon_{CD}\epsilon_{E'H'}\epsilon_{F'G'}\} + \epsilon_{AB}\Phi_{CDE'F'}\epsilon_{G'H'}$$
$$+ \epsilon_{CD}\Phi_{ABG'H'}\epsilon_{E'F'}, \tag{10·9}$$

where $$\Psi_{ABCD} = \Psi_{(ABCD)},\ \Phi_{ABC'D'} = \Phi_{(AB)(C'D')} = \bar{\Phi}_{C'D'AB}$$

and $$\Lambda = \tfrac{1}{24}R.$$

The spinor Ψ_{ABCD} corresponds to the Weyl conformal tensor similarly to (2·8) (i.e. put $\Lambda = 0$, $\Phi_{ABC'D'} = 0$ in (10·9) and replace $R_{\mu\nu\rho\tau}$ by $C_{\mu\nu\rho\tau}$) and $\Phi_{ABC'D'}$ is the spinor translation of $\frac{1}{8}Rg_{\mu\nu} - \frac{1}{2}R_{\mu\nu}$, i.e. of minus one-half the trace-free part of the Ricci tensor. Also

$$P_{ABC'D'} = P_{\mu\nu}\sigma^{\mu}_{AC'}\sigma^{\nu}_{BD'} = \Lambda\epsilon_{AB}\epsilon_{C'D'} - \Phi_{ABC'D'}$$

Hence, (10·5), (10·6) translate as follows:

$$\tilde{P}_{ABC'D'} = \Omega^2 P_{ABC'D'} - \Omega\nabla_{AC'}\nabla_{BD'}\Omega + \tfrac{1}{2}\epsilon_{AB}\epsilon_{C'D'}(\nabla_{EF'}\Omega)(\nabla^{EF'}\Omega), \tag{10·10}$$

$$\tilde{\Psi}_{ABCD} = \Omega^2\Psi_{ABCD}. \tag{10·11}$$

(Note that tilde quantities are translated by using $\tilde{\sigma}^{\mu}_{AB'}$, the others by using $\sigma^{\mu}_{AB'}$.) If desired (10·10) can be split into two equations:

$$\tilde{\Phi}^{C'D'}_{AB} = \Omega^2\Phi^{C'D'}_{AB} + \Omega\nabla^{C'}_{(A}\nabla^{D'}_{B)}\Omega, \tag{10·12}$$

$$\tilde{\Lambda} = \Omega^2\Lambda - \tfrac{1}{4}\Omega\nabla_{AB'}\nabla^{AB'}\Omega + \tfrac{1}{2}(\nabla_{AB'}\Omega)(\nabla^{AB'}\Omega). \tag{10·13}$$

11. The role of the cosmological constant

There is an important result concerning the boundary hypersurface $\mathscr{I}$, for an asymptotically simple space-time, which can be readily obtained at this stage. Assume, with Einstein, that $\tilde{R}_{\mu\nu} - \frac{1}{2}\tilde{R}\tilde{g}_{\mu\nu} + \lambda\tilde{g}_{\mu\nu}$ is a constant negative multiple of the energy-momentum tensor, λ being a (possibly zero) cosmological constant. Assume, also, that (in the neighbourhood of $\mathscr{I}$ at least) the energy and pressure densities‡ are non-negative definite, or, alternatively that the trace of the energy-momentum tensor vanishes near $\mathscr{I}$. Then, in the first case,

$$\{\Omega^2(\tilde{R}_{\mu\nu} - \tfrac{1}{2}\tilde{R}\tilde{g}_{\mu\nu} + \lambda\tilde{g}_{\mu\nu})\}x^{\mu}x^{\nu} \leqslant 0 \tag{11·1}$$

for any real x^{μ}. Now, it is clear from (10·5) that $\Omega^2\tilde{P}_{\mu\nu}$ (or, equivalently, from (10·10), that $\tilde{P}_{ABC'D'}$) remains finite and continuous at $\mathscr{I}$, its values there being defined by the right-hand side of (10·5). ($\tilde{P}_{\mu\nu}$, like $\tilde{g}_{\mu\nu}$, may be infinite on $\mathscr{I}$—the slight awkwardness involved here, would be avoided in the spinor formalism.) Hence, the expression inside the curly brackets in (11·1) remains finite and continuous at $\mathscr{I}$, whence (11·1) remains true on $\mathscr{I}$. From (10·5) we get

$$(\Omega^2\tilde{P}_{\mu\nu}) = ag_{\mu\nu} \quad on \quad \mathscr{I}, \tag{11·2}$$

† These definitions of Ψ_{ABCD}, $\Phi_{ABC'D'}$, Λ agree with those given by Newman & Penrose (1962) and they differ by simple factors from those given by Penrose (1960).

‡ The idea due to McCrae (1959) of simulating a cosmological constant by the introduction of negative pressures will not be used here.

where $$a = \tfrac{1}{2}(\nabla_\rho \Omega)\,(\nabla^\rho \Omega).$$

From (11·2), (10·4), $\tilde{R} = 24a$, $\Omega^2 \tilde{R}_{\mu\nu} = 6a g_{\mu\nu}$ on $\mathscr{I}$, so that substituting in (11·1) we get

$$(\lambda - 6a)\, g_{\mu\nu} x^\mu x^\nu \leqslant 0$$

on $\mathscr{I}$. This is violated either for timelike or spacelike vectors unless $\lambda = 6a$, i.e.

$$\lambda = 3(\nabla_\rho \Omega)\,(\nabla^\rho \Omega), \tag{11·3}$$

on $\mathscr{I}$. If, on the other hand, it is assumed that the trace of the energy-momentum tensor vanishes in the neighbourhood of $\mathscr{I}$, then (11·3) can be obtained at once from $\tilde{R} = 24a$ (compare Penrose 1963).

Because of condition 3 for an asymptotically simple space-time, the vector $\nabla_\rho \Omega$ is non-zero and normal to $\mathscr{I}$. Thus (11·3) gives us the following result: *the boundary hypersurface $\mathscr{I}$ for an asymptotically simple space-time is spacelike, timelike, or null according as the cosmological constant λ is positive, negative or zero*†. Note that in the Minkowski and Schwarzschild cases, $\mathscr{I}$ is null, in agreement with the above. Generally, a space-time will be considered *asymptotically flat* only if $\mathscr{I}$ is null. *Asymptotically de Sitter* space-times will have $\mathscr{I}$ spacelike (or timelike).

Equation (11·3) also has another significant consequence. Since the equation $\nabla_\rho \Omega \neq 0$ was not used in obtaining it, we have the following result: *if the cosmological constant λ is non-zero, then the condition $\nabla_\rho \Omega \neq 0$ at $\mathscr{I}$ can be deduced from the other conditions of asymptotic simplicity.*

12. Conformal behaviour of zero rest-mass fields

Let $\phi_{AB\ldots L}$ be symmetric with $2s(>0)$ indices satisfying the spin s zero rest-mass equation (2·5):

$$\nabla^{AM'} \phi_{AB\ldots K} = 0.$$

In order to preserve this under conformal transformation (compare McLennan 1956, Penrose 1962; see also Buchdahl 1959) the field $\phi_{A\ldots K}$ must be chosen to be a conformal density of weight $-s-1$:

$$\tilde{\phi}_{AB\ldots K} = \Omega^{s+1} \phi_{AB\ldots K}, \tag{12·1}$$

for then (10·8) gives, immediately, $\tilde{\nabla}^{AM'} \tilde{\phi}_{AB\ldots K} = 0$.

For spin zero, equation (2·12) must be used:

$$\Box \phi \equiv \{\nabla_{AM'} \nabla^{AM'} + 4\Lambda\}\, \phi = 0. \tag{12·2}$$

Then if $\tilde{\phi} = \Omega\phi$ (in accordance with (12·1)), it may be verified, with the help of (10·13) that $\tilde{\Box}\tilde{\phi} = 0$ as required. (Note that if $\tilde{\Lambda} = 0$, (10·13) can be written $\Box \Omega^{-1} = 0$. This is consistent with the above since Ω^{-1} may be thought of as a conformal density of weight -1, and $\tilde{\Box}\{\Omega\Omega^{-1}\} = 4\tilde{\Lambda} = 0$.)

In the special case of a Maxwell field, (12·1) becomes $\tilde{\phi}_{AB} = \Omega^2 \phi_{AB}$. Defining $F_{\mu\nu}$ according to (2·7) we get (via (10·7))

$$\tilde{F}_{\mu\nu} = F_{\mu\nu}.$$

† If $\lambda > 0$ this result requires only that the *energy* density be non-negative definite (x^μ timelike); if $\lambda < 0$ it requires only that the *pressures* be non-negative definite (x^μ spacelike). The same applies to the following result that $\nabla_\rho \Omega \neq 0$.

This simply expresses the well-known fact (Cunningham 1910; Bateman 1910) that Maxwell's free field equations are invariant under conformal transformation. In the case of spin 2, if we define $K_{\mu\nu\rho\sigma}$ according to (2·8) we get†

$$\tilde{K}_{\mu\nu\rho\sigma} = \Omega^{-1}K_{\mu\nu\rho\sigma}. \tag{12·3}$$

It was noted in §2 that, if Einstein's empty-space equations $\tilde{R}_{\mu\nu} = \lambda\tilde{g}_{\mu\nu}$ hold, the Weyl tensor (which equals the Riemann tensor if $\lambda = 0$) will satisfy the spin 2 zero rest-mass equation:

$$\tilde{\nabla}_{[\tau}\tilde{C}_{\mu\nu]\rho\sigma} = 0,$$

which is equivalent to

$$\tilde{\nabla}^{AM'}\tilde{\Psi}_{ABCD} = 0 \tag{12·4}$$

Thus, we can put

$$\tilde{K}_{\mu\nu\rho\sigma} = \tilde{C}_{\mu\nu\rho\sigma}, \quad \tilde{\phi}_{ABCD} = \tilde{\Psi}_{ABCD}. \tag{12·5}$$

It is important to observe, however, the difference between the conformal behaviour given by (10·6) and (10·11), and that given by (12·3) and the relevant (12·1). (This is related to the fact that $R_{\mu\nu}$ will generally *not* be proportional to $g_{\mu\nu}$ whereas $\tilde{R}_{\mu\nu}$ is.) Thus we must have

$$\Omega K_{\mu\nu\rho\sigma} = C_{\mu\nu\rho\sigma}, \quad \Omega\phi_{ABCD} = \Psi_{ABCD}. \tag{12·6}$$

This will be of great importance later: in empty space-time, the above $K_{\mu\nu\rho\sigma}$, ϕ_{ABCD} will satify the free-field zero rest-mass equations (2·9), (2·5) in $\mathscr{M}$.

13. The general peeling-off property

In order to obtain the peeling-off property of the p.n.d. (see §§1 and 6) for a zero rest-mass field in an asymptotically simple space-time, it will be valuable to be able to interpret the relevant concepts in a conformally invariant way. It is required to examine the spin s field $\phi_{AB\ldots K}$ along a null geodesic g, on which is defined an affine parameter r. To each point of g is assigned a spinor ξ_A which corresponds to the tangent (null) direction to g at that point. The normalization for ξ_A is to be determined up to a single complex multiplier by the condition that ξ_A be parallelly propagated along g.

Thus (according to the $g_{\mu\nu}$ metric),

$$\xi^B\bar{\xi}^{C'}\nabla_{BC'}\xi_A = 0, \tag{13·1}$$

since $\xi^B\bar{\xi}^{C'}\nabla_{BC'}(= \{\sigma^\mu_{BC'}\xi^B\bar{\xi}^{C'}\}\nabla_\mu)$ is an operator of propagation derivative along g. In order to preserve (13·1) under the conformal transformation (10·1) we put

$$\tilde{\xi}_A = \Omega^{\frac{1}{2}}\xi_A \tag{13·2}$$

(see 10·8). Note that the fact that g is a geodesic is contained in relation (13·1) since the tangent directions are parallelly propagated. Hence, the above conformal invariance for (13·1) also contains the well-known and important fact that null geodesics transform to null geodesics under conformal transformation.

† The direct tensor proof that (12·3) leads to conformal invariances for $\nabla_{[\tau}K_{\mu\nu]\rho\sigma} = 0$ is considerably more involved that the spinor argument. In its spinor form, the result was also independently obtained by Szekeres (1963).

The scaling of the parameter r can be fixed by

$$\xi^{B}\bar{\xi}^{C'}\nabla_{BC'}r = 1, \tag{13·3}$$

and this ensures that r is an affine parameter for the $g_{\mu\nu}$ metric. To preserve (13·3) we put $\mathrm{d}\tilde{r} = \Omega^{-2}\mathrm{d}r$, so that

$$\tilde{r} = \int \Omega^{-2}\mathrm{d}r \tag{13·4}$$

is the required affine parameter for the $\tilde{g}_{\mu\nu}$ metric.

It will be convenient also to introduce an auxiliary spinor η_A at each point of g subject to

$$\xi_A\eta^A = 1, \quad \xi^{B}\bar{\xi}^{C'}\nabla_{BC'}\eta_A = 0, \tag{13·5}$$

so that η^A is also parallelly propagated along g, and (see (10·8), (13·2)) to preserve these relations under conformal transformation by putting

$$\tilde{\eta}_A = \Omega^{-\frac{1}{2}}\eta_A + b\Omega^{\frac{1}{2}}\xi_A,$$

where $\xi^{B}\bar{\xi}^{C'}\nabla_{BC'}b = \Omega^{-2}\eta^{B}\bar{\xi}^{C'}\nabla_{BC'}\Omega$, that is (see (13·3)),

$$b = \int \Omega^{-2}\eta^{B}\bar{\xi}^{C'}\nabla_{BC'}\Omega\,\mathrm{d}r. \tag{13·6}$$

(Recall that the ξ_A, r and η_A of §5 were also constrained to satisfy (13·1), (13·3) and (13·5); see (5·12), (5·14) and (5·5).)

Suppose, now, that space-time $\tilde{\mathscr{M}}$ is asymptotically simple and related to $\mathscr{M}$ as in §9. The zero rest-mass field $\tilde{\phi}_{A\ldots K}$ in $\tilde{\mathscr{M}}$ will be called *asymptotically regular* if $\phi_{A\ldots K}$ exists throughout $\mathscr{M}$ which is related to $\tilde{\phi}_{A\ldots K}$ in the interior regions by (12·1) and which is *continuous* at the boundary $\mathscr{I}$. I shall show that *any asymptotically regular field satisfies the peeling-off property*.

The statement of the peeling-off property given in (6·1) must be suitably adapted so as to apply to asymptotically simple curved space-times. In order to give a meaning to '$O(\tilde{r}^{-1-k})$', it must be possible to compare expressions (6·1) at different points of g. This may be done by making the left-hand side a scalar by using the auxiliary spinor η_A. The peeling-off property is then

$$\tilde{\phi}_{A\ldots FG\ldots K}\tilde{\eta}^{A}\ldots\tilde{\eta}^{F}\tilde{\xi}^{G}\ldots\tilde{\xi}^{K} = O(\tilde{r}^{-k-1}) \tag{13·7}$$

along any null geodesic g, for $k = 0, 1, \ldots, 2s$, where $\tilde{\xi}$ appears k times and $\tilde{\eta}$ appears $2s-k$ times. Note that, given g, (13·7) is independent of the choice of $\tilde{\eta}^A$, $\tilde{\xi}^A$ and r satisfying (the tilde versions of) (13·5), (13·1) and (13·3) (with $\tilde{\xi}^A$ tangent to g). The freedom of choice is given by $\tilde{\eta}^A \to \alpha\tilde{\eta}^A + \beta\tilde{\xi}^A$, $\tilde{\xi}^A \to \alpha^{-1}\tilde{\xi}^A$, $\tilde{r} \to \alpha\bar{\alpha}\tilde{r} + \gamma$, where α, β, γ are constant along g and this manifestly does not affect (13·7) ($\alpha \neq 0$, $\gamma = \bar{\gamma}$). In flat space-time, (13·7) is clearly equivalent to (6·1).

Let the null geodesic g meet $\mathscr{I}$ in a point G. If x, y are quantities defined on g, I shall use the notation $x \sim y$ to mean the following: 'the *limit* of x/y exists at G and is non-zero there'. Clearly $x \sim y$ is the same as $y \sim x$. Now, from condition 3 for asymptotic simplicity we have $\nabla_{AB'}\Omega \neq 0$ on $\mathscr{I}$, whence (along g)

$$\mathrm{d}\Omega/\mathrm{d}r \equiv \xi^{A}\bar{\xi}^{B'}\nabla_{AB'}\Omega \neq 0 \quad \text{at} \quad G.$$

(The only occasions on which this deduction could fail would be if g were to *touch* $\mathscr{I}$ at G. This possibility will be ignored here as it could normally† only arise if $\mathscr{I}$ were timelike and 'concave' in places. Such a situation would seem unreasonable physically: g could re-enter the physical space-time after touching $\mathscr{I}$.) Also $\mathrm{d}\Omega/\mathrm{d}r$ exists by condition 3. Hence

$$\Omega \sim r - r_0, \tag{13·8}$$

r_0 being the value of r at G, so by (13·4)

$$\Omega \sim \tilde{r}^{-1}. \tag{13·9}$$

The normal to $\mathscr{I}$ at G and the tangent to g at G span a 2-plane element which contains one other null direction distinct from that of g. (Again, this could only fail if g were to touch $\mathscr{I}$.) Let us fix η_A by choosing it to correspond to this other null direction at G. Then $\nabla_{AB'}\Omega$ will be a linear combination of $\xi_A\bar{\xi}_{B'}$ and $\eta_A\bar{\eta}_{B'}$ at G, whence

$$\eta^A\bar{\xi}^{B'}\nabla_{AB'}\Omega = 0 \quad \text{at} \quad G.$$

From the C^2 differentiability of Ω at G, we obtain

$$\eta^A\bar{\xi}^{B'}\nabla_{AB'}\Omega = O(r - r_0),$$

which, with (13·6), (13·8) gives $\quad b = O(\ln\Omega)$,

so that

$$\tilde{\eta}_A = \Omega^{-\frac{1}{2}}\eta_A + \xi_A O(\Omega^{\frac{1}{2}}\ln\Omega). \tag{13·10}$$

(Note that this holds with the above choice of η_C *whichever* $\tilde{\eta}_A$ is chosen.)

Substituting (13·10), (13·2) and (12·1) into the left-hand side of (13·7), we obtain (near G)

$$\Omega^{k+1}\phi_{A\ldots FG\ldots K}\eta^A\ldots\eta^F\xi^G\ldots\xi^K + O(\Omega^{k+2}\ln\Omega). \tag{13·11}$$

The asymptotic regularity of $\phi_{A\ldots K}$ implies that it is bounded in the neighbourhood of G, as are η^A and ξ^A. But $\Omega^{k+1} \sim \tilde{r}^{-1-k}$ by (13·9), so that (13·11) is $O(\tilde{r}^{-k-1})$ as is required for (13·7). This establishes the peeling-off property.

Observe that only the *boundedness* of $\phi_{A\ldots K}$ in the neighbourhood of $\mathscr{I}$ has been used to establish (13·7). From *continuity* of $\phi_{A\ldots K}$ at $\mathscr{I}$ (asymptotic regularity) there follows a slightly stronger version‡ of the peeling-off property:

$$\lim_{\tilde{r}\to\pm\infty}(\tilde{r}^{k+1}\tilde{\phi}_{A\ldots FG\ldots K}\tilde{\eta}^A\ldots\tilde{\eta}^F\tilde{\xi}^G\ldots\tilde{\xi}^K) \textit{ exists} \tag{13·12}$$

($k\xi$'s and $2s-k\eta$'s). The value of this limit gives, of course, the relevant component of $\phi_{A\ldots K}$. Note that (13·12) is to hold both to the past and to the future along g, but the two limits need not be the same.

Of particular importance are the *radiation fields* given by (13·12) with $k = 0$. These are the advanced and retarded '$\tilde{r}^{-1}$ part' of the field $\tilde{\phi}_{A\ldots K}$. The *incoming* radiation field is given by (13·12) if the past limit is taken, the *outgoing* radiation field if the future limit is taken. Thus, the incoming and outgoing radiation fields are measured simply by

$$\phi = \phi_{A\ldots K}\eta^A\ldots\eta^K \tag{13·13}$$

† The only other situation in which g might touch $\mathscr{I}$, could sometimes arise if $\mathscr{I}$ were null in some places and non-null in others. However, this would be excluded by the physical assumptions made in § 11.

‡ This is, in fact, a more satisfactory statement of the peeling-off property than that given in (6·1) and (13·7). The argument given in § 6 could probably be suitably modified by re-defining the meaning of the order symbols and of the condition of asymptotic smoothness.

at the past and future end-points of g, respectively. (The choices of η^A at the past and future end-points may not together be consistent with (13·5), but this is unimportant.) Note that from (13·12),

$$\lim_{r\to\infty}(\tilde{r}\tilde{\phi}_{AB\ldots K}) = \phi\tilde{\xi}_A\tilde{\xi}_B\ldots\tilde{\xi}_K,$$

which illustrates the fact that the p.n.d. all coincide for the radiation part of the field. It also emphasizes that ϕ is not dependent on any particular choice of $\tilde{\eta}^A$, given g (and $\tilde{\xi}_A$).

One important difference emerges at this stage, between space-times in which $\mathscr{I}$ is null and those in which it is spacelike or timelike. If $\mathscr{I}$ is null, then the null direction determined by η^A is simply the normal (i.e. the null tangent) direction to $\mathscr{I}$ at G. Thus, η^A is essentially independent of the particular null geodesic g through G. The radiation field is effectively uniquely defined in an origin independent way. If, on the other hand, $\mathscr{I}$ is not null, then the null direction determined by η_A will change if g is changed to another null geodesic through G. In this case, the radiation field will become mixed up with the other components of the field when g is changed and so is *essentially* origin dependent. Thus, for an asymptotically simple space-time (satisfying the physical assumptions of § 11), we have the following important result: *the concept of the radiation field for a zero rest mass field can be given an origin independent meaning if and (for the case of positive spin) only if the cosmological constant vanishes.*

If $\mathscr{I}$ is null, other features also arise which do not occur if it is spacelike or timelike. For example, if the radiation field vanishes in some direction (as in fact it must—along at least $2s$ null geodesics through each event, some perhaps coincident) then the limiting component given by $k = 1$ in (13·12) ('$\tilde{r}^{-2}$ part') will be effectively uniquely determined in that direction in an origin independent way when $\mathscr{I}$ is null. For if (13·13) vanishes, then $\phi_{A\ldots JK}\eta^A\ldots\eta^J$ is proportional to η_K. Hence,

$$\phi_{A\ldots JK}\eta^A\ldots\eta^J\xi^K$$

is independent of $\xi^K(\xi^K\eta_K = -1)$. If the radiation field does *not* vanish, then this component of the field would be only origin independent up to the addition of a multiple of the radiation field. This type of situation clearly persists for the remaining components of $\phi_{A\ldots K}$.

More important, moreover, is the relationship between the radiation field and a form of *initial value problem* which can be applied when $\mathscr{I}$ is null. Expression (13·13) can be used as constraint-free initial data† for the field $\phi_{A\ldots K}$ on any suitable null

† An account of this by the author entitled 'Null hypersurface initial data for classical fields of arbitrary spin and for general relativity' has been widely circulated as a preprint, cf. Bergmann (1963). This is to be published later in an extended form. For a brief outline of the main ideas, see Komar (1962).

One point that emerges is that a quantity which is effectively the time derivative of the radiation field as defined here, is in some ways more significant than the radiation field itself. If the vanishing of *this* quantity on $\mathscr{I}^-$ (or $\mathscr{I}^+$) were to be interpreted as the absence of incoming (or outgoing) radiation, then this would be more in line with the Sommerfeld radiation condition, at least in the scalar case. (A static scalar monopole is not normally thought of as radiative!) For higher spin the relation is not so clear, however, particularly owing to the different connexions between energy flow and the radiation fields.

hypersurface—with which η^A is to be associated. When $\mathscr{I}$ is null, either $\mathscr{I}^-$ (past null infinity) or $\mathscr{I}^+$ (future null infinity) would form such a hypersurface. This implies that either the incoming or the outgoing radiation field (as defined here) can be used as the complete irredundant initial data for the field $\phi_{A\ldots K}$ if $\mathscr{I}$ is null.

When $\mathscr{I}$ is spacelike, the origin dependence of the radiation field leads to a remarkable phenomenon. The separation of fields into 'advanced' and 'retarded' parts is not in general possible ($s > \frac{1}{2}$) in such cases. For example, for a purely retarded field, it would be necessary for *all* components of $\phi_{A\ldots K}$ to vanish on $\mathscr{I}^-$. For most source distributions this is not possible, there being not enough initial data freedom available for $\phi_{A\ldots K}$. (This matter is discussed at greater length in Penrose (1964).)

14. Peeling-off in the Einstein–Maxwell theory

In §13, in order to obtain the peeling-off property in general, it was necessary to assume asymptotic regularity for all the zero rest-mass fields involved. (Note that the presence of interactions in no way affects the arguments of §13.) In this section it will be shown that if, instead, the appropriate field equations—which couple the relevant fields to physical curvature quantities—are assumed to hold in the neighbourhood of $\mathscr{I}$, then the assumption of asymptotic regularity can sometimes be dispensed with. Peeling-off will therefore follow merely from the asymptotic simplicity of the space-time.

In the first instance it will be shown that if Einstein's empty-space equations (with or without cosmological constant) hold (for $\tilde{g}_{\mu\nu}$) in the neighbourhood of $\mathscr{I}$, then the gravitational field satisfies the peeling-off property. Also, the condition $\nabla_\mu \Omega \neq 0$ can be deduced from the other conditions of asymptotic simplicity. Secondly, if the Einstein–Maxwell equations with positive cosmological constant hold in the neighbourhood of any point Q of $\mathscr{I}$, then both the gravitational and the electromagnetic fields satisfy the peeling-off property with respect to null geodesics through Q. Thirdly, if the Einstein–Maxwell equations without cosmological constant hold in the neighbourhood of $\mathscr{I}$, then both the gravitational and the electromagnetic fields satisfy the peeling-off property—except possibly for certain very special pathological cases which cannot be taken seriously from a physical point of view. These pathological cases, if they exist at all, would have the property that the electromagnetic field falls off faster than $\tilde{r}^{-1}$, with in fact† $\tilde{F}_{\mu\nu} = O(\tilde{r}^{-\frac{3}{2}})$. Furthermore, the total charge of the system would be necessarily zero. The unphysical nature of such a situation, if such exists, is brought out by the fact that the addition of a charge (or of electromagnetic radiation) to the system, no matter how small, would lead to an inconsistency with the asymptotic simplicity conditions. It should be mentioned that the arguments used here to obtain peeling-off in the presence of an electromagnetic field require that Ω be a C^6 function on $\mathscr{M}$, rather than just (say) a C^3 function.

† Instead of peeling-off, the electromagnetic field would behave as $\tilde{r}^{-\frac{3}{2}}, \tilde{r}^{-\frac{3}{2}}, \tilde{r}^{-\frac{3}{2}}$ and the gravitational field as $\tilde{r}^{-1}, \tilde{r}^{-2}, \tilde{r}^{-3}, \tilde{r}^{-3}, \tilde{r}^{-4}$, or else the electromagnetic field goes as $\tilde{r}^{-\frac{5}{2}}$ with peeling-off holding for the gravitational field.

The arguments in the case of non-vanishing cosmological constant λ turn out to be considerably easier than those for which $\lambda = 0$. When $\lambda \neq 0$ purely local arguments can be used. However, a partly global argument is necessary when $\lambda = 0$. According to §11, this is the case when $\mathscr{I}$ is null. It will therefore be necessary, before proceeding with calculations, to make some topological observations concerning $\mathscr{I}$, which will hold for *all* asymptotically simple space-times in which $\mathscr{I}$ is null.

If $\mathscr{I}$ is null, then it can be separated out into two disjoint parts $\mathscr{I}^-$ and $\mathscr{I}^+$. The points of $\mathscr{I}^-$ are distinguished from those of $\mathscr{I}^+$ by the fact that their future null cones, rather than their past null cones, lie in $\mathscr{M}$. Any null geodesic in $\mathscr{M}$ not on $\mathscr{I}$, contains a past endpoint on $\mathscr{I}^-$ and a future endpoint on $\mathscr{I}^+$. (This much is true also if $\mathscr{I}$ is spacelike.) Furthermore, since $\mathscr{I}^-$ and $\mathscr{I}^+$ are null and non-singular, they are generated by a double infinity of non-intersecting null geodesics, one through each point. Finally, it also turns out that the topology of each of $\mathscr{I}^-$ and $\mathscr{I}^+$ is necessarily $S^2 \times E^1$—that is, a three-dimensional cylinder having cross-sections which are topological spheres (S^2's). The generators (E^1's) of the cylinders are the null geodesics lying on $\mathscr{I}$. An argument which establishes this topological structure for $\mathscr{I}$ is outlined in an appendix. The reader, may, however be willing simply to *assume* this structure as being desirable, from a *physical* point of view.

The following lemma will be needed to deal with the case when $\mathscr{I}$ is null. It depends essentially on the above-mentioned topological structure for $\mathscr{I}$.

LEMMA I. *Let $\mathscr{I}$ be null. Let the null direction in $\mathscr{I}$ be represented by a spinor ι_A satisfying*

$$\iota^A \nabla_{A(B'} \{\bar{\iota}_{D)} \iota_{C\}} = 0 \tag{14·1}$$

on some smooth space-like cross-section $\mathscr{X}$ of $\mathscr{I}^{\pm}$. Suppose a symmetric continuous spinor field $\chi_{AB\ldots K}$ satisfies

$$\iota^A \chi_{AB\ldots K} = 0, \quad \iota^L \nabla_{LM'} \chi_{AB\ldots K} = 0 \tag{14·2}$$

on $\mathscr{X}$. Then $\chi_{A\ldots K} = 0$ on $\mathscr{X}$.

Proof. Note, first, that a substitution $\iota_A \to e^{i\theta} \iota_A$ (θ real—not necessarily continuous) does not affect the conditions of the lemma. Now $\mathscr{X}$, being a smooth space-like S^2 can be transformed under a conformal map $g^*_{\mu\nu} = \Theta^{-2} g_{\mu\nu}$ into (intrinsically) an infinite Euclidean metric plane $\mathscr{X}^*$, except that the one point X, of $\mathscr{X}$ given by $\Theta = 0$, is mapped to infinity in $\mathscr{X}^*$. (To achieve this, $\mathscr{X}$ may, if desired, be first mapped conformally into a *metric sphere* and then into the metric plane $\mathscr{X}^*$ by means of the transformation of a stereographic projection from X.) The equations (14·1) and (14·2) are preserved under this conformal mapping if (see (10·8))

$$\iota^*_A = \iota_A, \quad \chi^*_{AB\ldots K} = \Theta^{\frac{1}{2}k} \chi_{AB\ldots K}, \tag{14·3}$$

where k is the number of indices of $\chi_{A\ldots K}$.

Let $\mathscr{X}^*$ be the Argand plane of a complex parameter $\sqrt{2}\zeta$. Let the complex vector $m^{*\mu}$ be the gradient of $-\zeta$ in the plane $\mathscr{X}^*$, so that if α is defined on $\mathscr{X}^*$ we have

$$\frac{\partial \alpha}{\partial \zeta} = \delta \alpha, \quad \frac{\partial \alpha}{\partial \bar{\zeta}} = \delta \alpha \tag{14·4}$$

where $$\delta \equiv m^{*\mu}\nabla^*_\mu, \quad \bar{\delta} \equiv \overline{m}^{*\mu}\nabla^*_\mu \tag{14·5}$$

with $m^*_\mu \overline{m}^{*\mu} = -1$, $m^*_\mu m^{*\mu} = 0$. Since m^*_μ is tangent to $\mathscr{X}^*$, it is orthogonal to the null direction represented by ι^*_A. (The null direction in $\mathscr{I}^{\pm}$, being the *normal* to $\mathscr{I}^{\pm}$, is orthogonal to $\mathscr{X}$.) Hence, $m^*_{AB'} = \sigma^{*\mu}_{AB'} m^*_\mu = \Theta\sigma^\mu_{AB'} m^*_\mu$ has the form

$$m^*_{AB'} = \iota_A \bar{\xi}_{B'}, \tag{14·6}$$

where, by the normalization of m^*_μ, $|\iota_A \xi^A| = 1$, or else it is $\overline{m}^*_{AB'}$ that has the above form. Choose the orientation of the Argand plane so that (14·6), in fact, holds and use the freedom $\iota_A \to e^{i\theta}\iota_A (\xi_A \to e^{i\theta}\xi_A)$ to obtain

$$\iota_A \xi^A = 1. \tag{14·7}$$

Now, we have $\overline{m}^{*\mu}\delta m^*_\mu = 0$ since the vector m^*_μ is intrinsically constant for the plane $\mathscr{X}^*$. That is,

$$\xi^A \bar{\iota}^{B'} \delta\{\iota_A \bar{\xi}_{B'}\} = 0.$$

But from the asterisk version of (14·1) (using (14·5), (14·6)),

$$\xi^A \bar{\xi}^{B'} \delta\{\iota_A \bar{\iota}_{B'}\} = 0.$$

Subtracting these two relations and using (14·7) and its derivative, we get

$$\xi^A \delta\iota_A = 0. \tag{14·8}$$

From the first of conditions (14·2) and the symmetry of $\chi_{A\ldots K}$, we get

$$\chi^*_{AB\ldots K} = \alpha\iota_A\iota_B\ldots\iota_K \tag{14·9}$$

on $\mathscr{X}^*$ and from the asterisk version of the second, we get

$$\xi^A\xi^B\ldots\xi^K\delta\chi^*_{AB\ldots K} = 0. \tag{14·10}$$

Substituting (14·9) in (14·10) and using (14·8), (14·4), we see that α must be an analytic function of ζ on the whole of $\mathscr{X}$. Hence either $\alpha = 0$ or $\overline{\lim_{\zeta\to\infty}}|\alpha| > 0$. In the latter case it would follow from (14·3) that $\chi_{A\ldots K}$ is unbounded in the neighbourhood of X, violating continuity. Thus, $\chi_{A\ldots K} = 0$ on $\mathscr{X}$.

The next lemma applies whether or not $\mathscr{I}$ is null.

LEMMA II. *Let $\chi_{A\ldots K}$ be symmetric and defined throughout $\mathscr{M}$, with continuous derivative at $\mathscr{I}$. Suppose $\theta^{P'}_{B\ldots K}$ is continuous at $\mathscr{I}$ where $\theta^{P'}_{B\ldots K} = \nabla^{AP'}\{\Omega^{-1}\chi_{AB\ldots K}\}$ in the interior of $\mathscr{M}$. If $\mathscr{I}$ is null, suppose $\nabla^A_{(B'}\nabla^C_{D')}\Omega = 0$ on $\mathscr{I}$. Then $\Omega^{-1}\chi_{A\ldots K}$ is continuous at $\mathscr{I}$.*

Proof.

$$\Omega\theta^{P'}_{B\ldots K} = \nabla^{AP'}\chi_{AB\ldots K} - \Omega^{-1}\chi_{AB\ldots K}\nabla^{AP'}\Omega \tag{14·11}$$

is continuous and vanishes at $\mathscr{I}$. Hence $\chi_{AB\ldots K}\nabla^{AP'}\Omega = 0$ on $\mathscr{I}$. But $\nabla^{AP'}\Omega$ is non-zero and normal to $\mathscr{I}$. Therefore the matrix $\nabla^{AP'}\Omega$ is non-singular if $\mathscr{I}$ is *not* null. In this case we must have $\chi_{AB\ldots K} = 0$ on $\mathscr{I}$ and from the differentiability of $\chi_{A\ldots K}$ at $\mathscr{I}$, it then follows that $\Omega^{-1}\chi_{A\ldots K}$ is continuous at $\mathscr{I}$. If $\mathscr{I}$ *is* null then

$$\nabla^{AB'}\Omega = \pm\iota^A\bar{\iota}^B$$

at $\mathscr{I}^{\mp}$ for some ι^A, whence $\chi_{AB\ldots K}\iota^A = 0$ on $\mathscr{I}$. Multiply (14·11) by Ω:

$$\Omega\nabla^{AP'}\chi_{AB\ldots K} - \chi_{AB\ldots K}\nabla^{AP'}\Omega. \tag{14·12}$$

This must have vanishing derivative at $\mathscr{I}$. (For if $\alpha\ldots$ is continuous at $\mathscr{I}$, then $\Omega\alpha\ldots$ is differentiable at $\mathscr{I}$ with gradient $\alpha\ldots\nabla_\mu\Omega$.) Differentiating (14·12), we get

$$\nabla^{AP'}\chi_{AB\ldots K}\nabla^{LQ'}\Omega - \nabla^{LQ'}\chi_{AB\ldots K}\nabla^{AP'}\Omega - \chi_{AB\ldots K}\nabla^{LQ'}\nabla^{AP'}\Omega = 0 \tag{14·13}$$

on $\mathscr{I}$. With $\nabla^L_{(Q'}\nabla^A_{P')}\Omega = 0$ and $\nabla^{AP'}\Omega = \pm\iota^A\bar{\iota}^{P'}$ on $\mathscr{I}$, then (14·13) gives, after symmetrization over P', Q',

$$2\iota^{[L}\bar{\iota}_{(Q'}\nabla^{A]}_{P')}\chi_{AB\ldots K} \equiv \epsilon^{LA}\iota_M\bar{\iota}_{(Q'}\nabla^M_{P')}\chi_{AB\ldots K} = 0$$

on $\mathscr{I}$. The $\bar{\iota}_{Q'}$ may be removed from this equation (multiply by $\theta^{Q'}\theta^{P'}$, where $\theta^{Q'}$ is arbitrary, divide by $\theta^{Q'}\bar{\iota}_{Q'}$ and cancel out the arbitrary $\theta^{P'}$), whence

$$\iota_M\nabla^M_{P'}\chi_{AB\ldots K} = 0$$

on $\mathscr{I}$. Lemma I then gives $\chi_{A\ldots K} = 0$ on $\mathscr{I}$, since

$$\iota^A\nabla_{A(B'}\{\iota^C\bar{\iota}_{D')}\} = \pm\iota^A\nabla_{A(B'}\nabla^C_{D')}\Omega = 0$$

on $\mathscr{I}^{\mp}$ ($\iota^A\nabla_{AB'}$ operates tangentially in $\mathscr{I}$.) Hence $\Omega^{-1}\chi_{A\ldots K}$ is continuous, as required.

We are now in a position to discuss Einstein's empty-space equations. It is required to obtain the peeling-off property for the gravitational field. The result is almost immediate from lemma II. For the conformal curvature spinor Ψ_{ABCD} of $\mathscr{M}$ is C^1 at $\mathscr{I}$ and satisfies

$$\nabla^{AP'}\{\Omega^{-1}\Psi_{ABCD}\} = 0 \tag{14·14}$$

when the empty-space conditions hold (for $\tilde{\mathscr{M}}$) near $\mathscr{I}$, since by (12·6) if

$$\phi_{ABCD} = \Omega^{-1}\Psi_{ABCD}, \tag{14·15}$$

then ϕ_{ABCD} satisfies the spin 2 zero rest-mass equation (2·5). For the purposes of the present discussion the transformed *gravitational field* must be represented as this spin 2 field (14·15) (the untransformed field being represented by the physical Weyl spinor $\tilde{\Psi}_{ABCD}$—see (12·5)). If a cosmological term is present, then (by (11·3)) $\mathscr{I}$ is not null. In this case, lemma II tells us at once that ϕ_{ABCD}, given by (14·15), is continuous at $\mathscr{I}$. Thus the gravitational field is asymptotically regular, so by §13, the peeling-off property holds. If, on the other hand, the cosmological constant vanishes, then $\mathscr{I}$ is null so we must examine $\nabla^A_{(B'}\nabla^C_{D')}\Omega$. Now $\tilde{\Phi}_{ABC'D'}$ vanishes in empty space-time, whence by (10·12),

$$\nabla^A_{(B'}\nabla^C_{D')}\Omega + \Omega\Phi^{AC}_{B'D'} = 0. \tag{14·16}$$

This gives $\nabla^A_{(B'}\nabla^C_{D')}\Omega = 0$, on $\mathscr{I}$, so lemma II again applies and the peeling-off property holds†.

† In a preliminary account (Penrose 1963), instead of using (14·14), the equivalent, but more complicated method of applying Ricci identities to derivatives of Ω, was suggested.

Certain small errors in that account might profitably be pointed out here. On p. 67, column 2, line 21, '$C_{\mu\nu\rho\sigma}{}^{;\sigma} = 0$' should be '$C_{\mu\nu\rho\sigma}{}^{;\tau}m_\tau = 0$'; line 30, for consistency with the definition of N, we should have '$\Psi_4 = K_{\mu\nu\rho\sigma}n^\mu\overline{m}^\nu n^\rho\overline{m}^\sigma$'; line 40, we should have '$\tilde{l}^\mu = \tilde{r}^{-2}l^\mu, \tilde{m}^\mu = \tilde{r}^{-1}m^\mu$'; line 45, '$O(r^{-5-i})$' should be '$O(r^{-5+i})$'. On p. 68, the factor before the integral sign should be '$2^{-\frac{5}{2}}\pi^{-1}$', or, alternatively, '$\frac{1}{4}\pi^{-1}$' if the condition '$R = 12$' near the bottom of p. 67 is replaced by '$R = 6$' —a choice which makes the sections of $\mathscr{I}^+$ unit spheres.

Equation (14·16) has a further consequence. It tells us that if space-time is empty near $\mathscr{I}$, then the condition $\nabla_\mu\Omega \neq 0$ follows from the other conditions of asymptotic simplicity. For let g be a null geodesic meeting $\mathscr{I}$ at a point G. Let ξ_A and r be associated with g as in §13 (see (13·1), (13·3)). Multiplying (14·16) by $\xi_A\bar{\xi}^{B'}\xi_C\bar{\xi}^{D'}$, we get

$$\mathrm{d}^2\Omega/\mathrm{d}r^2 + \Omega\Phi = 0 \tag{14·17}$$

along g, where

$$\Phi = \Phi_{B'D'}^{AC}\xi_A\xi_C\bar{\xi}^{B'}\bar{\xi}^{D'}$$

is a continuous function of r. If $\nabla_\mu\Omega = 0$ at G, then $\mathrm{d}\Omega/\mathrm{d}r = 0$ at G. But it is not hard to see that the only solution of (14·17) for which $\Omega = 0 = \mathrm{d}\Omega/\mathrm{d}r$ at G is $\Omega \equiv 0$ on g, so the result follows.

In order to incorporate an electromagnetic field into the discussion another lemma is needed.

LEMMA III. *Let $\chi_{ABC'D'}$ be defined throughout $\mathscr{M}$ with continuous derivative at $\mathscr{I}$. Suppose ϕ_{AB} exists in the interior of $\mathscr{M}$ which is symmetric and C^1 differentiable and where $\chi_{ABC'D'} = \Omega^{2q}\phi_{AB}\bar{\phi}_{C'D'}$ in the interior of $\mathscr{M}$ near $\mathscr{I}$, with $\Omega^p\nabla^{AP'}\phi_{AB}$ bounded near $\mathscr{I}$, for some p, q; $p-1 < q \neq 0$. Then if $\mathscr{I}$ is spacelike, $\Omega^{-1}\chi_{ABC'D'}$ is continuous at $\mathscr{I}$. If $\mathscr{I}$ is null, corresponding to ι^A, then $\iota^A\chi_{ABC'D'} = 0$ on $\mathscr{I}$. If $\mathscr{I}$ is null with $\nabla^A_{(B'}\nabla^C_{D')}\Omega = a\iota^A\bar{\iota}_{B'}\iota^C\bar{\iota}_{D'}$ on $\mathscr{I}$ and $p < q = 1$, then $\Omega^{-1}\chi_{ABC'D'}$ is continuous at $\mathscr{I}$.*

Proof. Since $\chi_{ABC'D'}$ is continuous, it is of the form

$$\chi_{ABC'D'} = \psi_{AB}\bar{\psi}_{C'D'} \tag{14·18}$$

at $\mathscr{I}$ also. It is required to choose the phase of ψ_{AB} in such a way that ψ_{AB} is continuous and (except where $\chi_{ABC'D'} = 0$ on $\mathscr{I}$) C^1 differentiable. In the interior of $\mathscr{M}$, this would be achieved if $\psi_{AB} = \Omega^q\phi_{AB}$, but it is not clear that this would necessarily also give continuity at $\mathscr{I}$, since the phase of ϕ_{AB} might oscillate infinitely many times near $\mathscr{I}$. However, any obstructions to choosing a continuous ψ_{AB} generally, would be of a topological nature. We can avoid both difficulties by choosing $\psi_{AB} = \Omega^q\phi_{AB}$ on a hypersurface $\mathscr{I}^*$ in the interior of $\mathscr{M}$, which lies sufficiently close to $\mathscr{I}$ and then allowing ψ_{AB} to change smoothly from $\mathscr{I}^*$ to $\mathscr{I}$. We can also ensure that ψ_{AB} is C^1 at $\mathscr{I}$, wherever $\chi_{ABC'D'} \neq 0$ at $\mathscr{I}$. To avoid diffiiculties concerning the possibility that at some regions of $\mathscr{I}$, $\chi_{ABC'D'} = 0$ (and $\chi_{ABC'D'} \neq O(\Omega^2)$), while at others $\chi_{ABC'D'} \neq 0$, the following discussion will be concerned with the latter regions in the first instance. It is required to prove $\chi_{ABC'D'} = 0$ on $\mathscr{I}$ in any case.

We have

$$\phi_{AB} = \mathrm{e}^{\mathrm{i}\theta}\Omega^{-q}\psi_{AB},$$

where θ is real and differentiable in the interior of $\mathscr{M}$. Now,

$$\mathrm{e}^{-\mathrm{i}\theta}\Omega^{q+1}\nabla^{AP'}\phi_{AB} \equiv \Omega\nabla^{AP'}\psi_{AB} + \mathrm{i}\Omega\psi_{AB}\nabla^{AP'}\theta - q\psi_{AB}\nabla^{AP'}\Omega \to 0 \tag{14·19}$$

at $\mathscr{I}$, where $\nabla^{AP'}\theta$ might be unbounded near $\mathscr{I}$. Multiply (14·19) by $\bar{\psi}_{P'Q'}$ and take the part Hermitian in BQ'. This tells us that

$$\psi_{BA}(\nabla^{AP'}\Omega)\bar{\psi}_{P'Q'} = 0$$

at $\mathscr{I}$. This can occur only if the matrix ψ_{AB} is singular and therefore of the form $\psi_{AB} = \eta_A\eta_B$, at $\mathscr{I}$, in which case

$$\eta_A\bar{\eta}_{P'}\nabla^{AP'}\Omega = 0.$$

This tells us that the null vector represented by $\eta_A\bar{\eta}_{P'}$ is orthogonal to the normal to $\mathscr{I}$. If $\mathscr{I}$ is spacelike, this means $\eta_A = 0$, whence by (14·18), $\chi_{ABC'D'} = 0$. From differentiability, $\Omega^{-1}\chi_{ABC'D'}$ is continuous at $\mathscr{I}$, as required.

If $\mathscr{I}$ is null, we get $\iota^A\eta_A = 0$, so by (14·18), $\iota^A\chi_{ABC'D'} = 0$ on $\mathscr{I}$. Now the quantity $\psi_{AB}\nabla^{AP'}\Omega$ vanishes on $\mathscr{I}$, so it is of the form $\Omega\times$ a continuous expression. But if $q > p$, (14·19) is of the form $\Omega\times$ a continuous expression vanishing at $\mathscr{I}$. It follows that $\psi_{AB}\nabla^{AP'}\theta$ must be continuous at $\mathscr{I}$. Differentiating (14·19) and taking $q = 1$, we obtain

$$\{\nabla^{AP'}\psi_{AB} + \mathrm{i}\psi_{AB}\nabla^{AP'}\theta\}\nabla^{LQ'}\Omega - \nabla^{LQ'}\psi_{AB}\nabla^{AP'}\Omega - \psi_{AB}\nabla^{LQ'}\nabla^{AP'}\Omega = 0$$

at $\mathscr{I}$. If $\nabla^A_{(B'}\nabla^C_{D')}\Omega = a\iota^A\bar{\iota}_{B'}\iota^C\bar{\iota}_{D'}$, then the above reduces to

$$\iota_M\nabla^M_{P'}\{\mathrm{e}^{\mathrm{i}\theta}\psi_{AB}\} \to 0 \tag{14·20}$$

at $\mathscr{I}$. From the continuity at $\mathscr{I}$ of $\psi_{AB}\nabla^{AP'}\theta$ it follows that $\iota_A\nabla^{AP'}\theta$ is well-defined on $\mathscr{I}$. The operator $\iota_A\nabla^{AP'}$ determines differentiation within the null hypersurface $\mathscr{I}$. Hence θ is well-defined on $\mathscr{I}$ up to an additive constant and is differentiable within $\mathscr{I}$. With $\chi_{AB} = \mathrm{e}^{\mathrm{i}\theta}\psi_{AB}$ on $\mathscr{I}$ and the limit of (14·20) on $\mathscr{I}$ being taken, the proof is completed essentially as in lemma II. (Regions of ϕ at which $\chi_{ABC'D'}$ vanishes initially do not cause trouble since we simply put $\chi_{AB} = 0$ there.)

The result of lemma III has been stated in a somewhat more general form than is absolutely necessary for the application to the Einstein–Maxwell theory. It is possible that this might be of value if fields other than just gravitational and electromagnetic (e.g. neutrino) are present near $\mathscr{I}$. In fact lemma III can be generalized so as to apply to $\phi_{A\ldots K}$ of arbitrary positive spin (although it is doubtful if this could be used to obtain neutrino peeling-off). In this form, lemma II is effectively a special case.

One point that is important to observe, moreover, is that when $\mathscr{I}$ is spacelike, the arguments for lemmas II and III will apply *locally*—that is, in the neighbourhood of a point Q of $\mathscr{I}$. For, unlike the case of a null (or timelike) $\mathscr{I}$, we must expect that *finite* rest-mass fields are significant at some regions of $\mathscr{I}$, when $\mathscr{I}$ is spacelike. Timelike geodesics can meet $\mathscr{I}$ when, and only when, it is spacelike. (Particle-like bounded world-tubes may be thought of as represented on $\mathscr{M}$, when ϕ is spacelike, by regions meeting $\mathscr{I}$ at a set of discrete points. However, in this case, these points would normally be singular for $\mathscr{M}$ and not properly parts of $\mathscr{M}$.)

For the case of the Einstein–Maxwell equations, with or without cosmological term, we have (Penrose 1960)

$$\tilde{\Phi}_{ABC'D'} = k\tilde{\phi}_{AB}\bar{\tilde{\phi}}_{C'D'} \tag{14·21}$$

in the interior of $\mathscr{M}$ near $\mathscr{I}$, where k is a positive constant and where the (C^1) electromagnetic field $\tilde{\phi}_{AB}$ satisfies the spin 1 zero rest-mass equation according to $\tilde{\nabla}$. Hence (see (12·1)), if

$$\phi_{AB} = \Omega^{-2}\tilde{\phi}_{AB}, \tag{14·22}$$

then

$$\nabla^{AP'}\phi_{AB} = 0.$$

By (10·12), (14·21) and (14·22), we have

$$\nabla^A_{(B'}\nabla^C_{D')}\Omega + \Omega\Phi^{AC}_{B'D'} = k\Omega^3\phi^{AC}\overline{\phi}_{B'D'} \tag{14·23}$$

in the interior of $\mathscr{M}$. Denoting the left-hand side of (14·23) by $k\eta^{AC}_{B'D'}$ we see that $\eta_{ABC'D'}$ satisfies the conditions for $\chi_{ABC'D'}$ in lemma III with $q = \frac{3}{2}$. If the cosmological constant λ is positive, then $\mathscr{I}$ is spacelike and lemma III tells us that $\Omega^{-1}\eta^{AC}_{B'D'}$ is continuous at $\mathscr{I}$ (and is C^3, taking Ω to be a C^6 function and $\sigma^{\mu}_{AB'}$ a C^5 function at $\mathscr{I}$). Repeating the argument with $\Omega^{-1}\eta^{AC}_{B'D'}$ as $\chi^{AC}_{B'D'}$ and $q = 1$, it follows that $\Omega^{-2}\eta^{AC}_{B'D'}$ is continuous (and C^2) at $\mathscr{I}$. Repeating once more using $\Omega^{-2}\eta^{AC}_{B'D'}$ and $q = \frac{1}{2}$, we get $\Omega^{-3}\eta^{AC}_{B'D'}$ continuous C^1 at $\mathscr{I}$. Hence (14·23) tells us that the field ϕ_{AB} is bounded at $\mathscr{I}$. It follows from §13 that the electromagnetic field satisfies the peeling-off property.†

If $\lambda = 0$, then $\mathscr{I}$ is null and we have merely $\iota^A\eta_{ABC'D'} = 0$ on $\mathscr{I}$, where

$$\nabla_{AB'}\Omega = \pm\iota_A\bar{\iota}_{B'}$$

at $\mathscr{I}^{\mp}$. This means $\eta_{ABC'D'}$ is of the form $ak^{-1}\iota_A\iota_B\bar{\iota}_{C'}\bar{\iota}_{D'}$ on $\mathscr{I}$, whence (see (14·23),) $\nabla^A_{(B'}\nabla^C_{D')}\Omega$ has the form $a\iota^A\bar{\iota}_{B'}\iota^C\bar{\iota}_{D'}$. Unfortunately, however, lemma III will not fully apply now since $q = 1$ only at the second step. We obtain merely

$$\iota^A\psi_{AB} = 0, \quad \iota_M\nabla^M_{P'}\{e^{i\theta}\psi_{LB}\} = (q-1)e^{i\theta}\iota^A\nabla_{LP'}\psi_{AB} \tag{14·24}$$

at $\mathscr{I}$, where $\psi_{AB} = \Omega^q e^{-i\theta}\phi_{AB}$ is C^1. The possibility mentioned at the beginning of this section, of pathological situations in which peeling-off does not hold for the electromagnetic field, must now be considered.

Let g be a null geodesic meeting $\mathscr{I}$ at G, and associated with ξ_A. Then, taking the left-hand side $k\eta^{AC}_{B'D'}$ of (14·23) as C^3 (actually C^4) at $\mathscr{I}$, we have

$$\Omega^3\phi_{AB}\bar{\phi}_{C'D'} = \eta_{ABC'D'} = \alpha_{ABC'D'} + \Omega\beta_{ABC'D'} + \Omega^2\gamma_{ABC'D'} + \Omega^3\Sigma_{ABC'D'} \tag{14·25}$$

where $\alpha_{....}$, $\beta_{....}$, $\gamma_{....}$ are constant (parallelly propagated) along g and where $\Sigma_{....}$ is continuous at $\mathscr{I}$. This fixes the definitions of these quantities uniquely in terms of $\eta_{ABC'D'}$. If $\alpha_{....} \neq 0$, then we have $\eta_{ABC'D'} = \psi_{AB}\bar{\psi}_{C'D'}$ with ψ_{AB} a C^2 (actually C^4) function along g, so that

$$\psi_{AB} = \lambda_{AB} + \Omega\mu_{AB} + \Omega^2\Upsilon_{AB}, \tag{14·26}$$

with $\lambda_{..}$, $\mu_{..}$ constant along g and $\Upsilon_{..}$ continuous at $\mathscr{I}$. We have $\alpha_{....} = \lambda_{..}\bar{\lambda}_{..}$, $\beta_{....} = \lambda_{..}\bar{\mu}_{..} + \mu_{..}\bar{\lambda}_{..}$, etc. This gives ϕ_{AB} the form

$$\phi_{AB} = e^{i\theta}(\Omega^{-\frac{3}{2}}\lambda_{AB} + \Omega^{-\frac{1}{2}}\mu_{AB} + \Omega^{\frac{1}{2}}\Upsilon_{AB}), \tag{14·27}$$

where θ may vary along g. If, on the other hand, $\alpha_{....} = 0$ but $\beta_{....} \neq 0$, then $\eta_{ABC'D'}$ is of the form $\Omega\psi_{AB}\bar{\psi}_{C'D'}$ with ψ_{AB} a C^2 function. Hence, ϕ_{AB} is now of the form

$$\phi_{AB} = e^{i\theta}(\Omega^{-1}\lambda_{AB} + \mu_{AB} + \Omega\Upsilon_{AB}) \tag{14·28}$$

(where $\beta_{....} = \lambda_{..}\bar{\lambda}_{..}$, etc.) If $\alpha_{....} = \beta_{....} = 0 \neq \gamma_{....}$, then the form ϕ_{AB} is as in (14·27) again but with $\lambda_{AB} = 0$. The importance of all this is that we get two quite separate situations, namely those involving fractional powers of Ω and those with integral powers. If we attempt to modify a field such as (14·27) by adding to it, say, a small multiple of (14·28), we get an immediate contradiction with asymptotic simplicity, since the cross-terms would lead to fractional powers in (14·25) which would spoil the differentiability of $\eta_{ABC'D'}$.

† Since only boundedness and not continuity has been established for ϕ_{AB}, this is the slightly weaker version of this property. The stronger version probably also holds.

Now suppose $\alpha_{\dots} \neq 0$ and apply (14·24) to (14·26) with $q = \frac{3}{2}$. This gives

$$\lambda_{AB}\iota^A = 0, \quad \mu_{AB}\iota^A\iota^B = 0, \tag{14·29}$$

the second of these equations following since by (14·26)

$$\iota^A\iota^B\xi^L\xi^{P'}\nabla_{LP'}\psi_{AB} = \pm|\xi_Y\iota^Y|^2\mu_{AB}\iota^A\iota^B$$

at $\mathscr{I}$ and the left-hand side vanishes by (14·24), since

$$\iota_M\iota^B\nabla^M_{P'}\psi_{LB} = \iota_M\nabla^M_{P'}(\iota^B\psi_{LB}) = 0.$$

at $\mathscr{I}$. ($\iota_M\nabla^M_{P'}$ acts tangentially in $\mathscr{I}$ and $\bar{\iota}^{Y'}\iota_B\iota_M\nabla^M_{P'}\iota^B = \iota_B\iota_M\nabla^{(M}_{P'}\{\iota^{B)}\bar{\iota}^{Y'}\} = 0$.) Applying the method of §13 to (14·27) using (14·29), we then obtain the result that all components of the electromagnetic field fall off as $\tilde{r}^{-\frac{3}{2}}$ (or perhaps faster). 'Radiation field' (as defined here) would be absent. To obtain the total charge of the system, the component $\phi_{AB}\iota^A\xi^B$ may be integrated around a cross-section of $\mathscr{I}$ (or near $\mathscr{I}$). Because of (14·29), the first term that could contribute from (14·27) is $e^{i\theta}\mu_{AB}\Omega^{-\frac{1}{2}}$. However, by (14·24), (14·26),

$$\xi^L\xi^B\iota_M\bar{\xi}^{P'}\nabla^M_{P'}\{e^{i\theta}\lambda_{LB}\} = \pm\tfrac{1}{2}e^{i\theta}\mu_{AB}\iota^A\xi^B|\xi_X\iota^X|^2,$$

from which it follows, by a form of Stokes's theorem, that this integral vanishes.

It seems clear, therefore, that the physically meaningful fields ϕ_{AB} are of the form (14·28), not (14·27). Assuming the form (14·28) holds, then $\alpha_{\dots} = 0$ and application of lemma III with $q = 1$ gives $\beta_{\dots} = 0$. Hence $\lambda_{AB} = 0$ in (14·28). Thus ϕ_{AB} is bounded at $\mathscr{I}$ and the peeling-off property holds, as required.

Once peeling-off for the electromagnetic field has been thus established, it is a simple matter to obtain it for the gravitational field also. In place of (12·4), the Bianchi identity now gives †

$$\tilde{\nabla}^{AP'}\tilde{\Psi}_{ABCD} = -k\tilde{\phi}^{P'Q'}\tilde{\nabla}_{BQ'}\tilde{\phi}_{CD}$$

(Penrose 1960). In terms of the 'unphysical' quantities this becomes ((10·8), (14·22))

$$\begin{aligned}\nabla^{AP'}\phi_{ABCD} &= -\Omega k\bar{\phi}^{P'Q'}\nabla_{BQ'}\phi_{CD} - 3k\bar{\phi}^{P'Q'}\phi_{(CD}\nabla_{B)Q'}\Omega \\ &= -\Omega k\nabla_{BQ'}\{\phi_{CD}\bar{\phi}^{P'Q'}\} - 3k\bar{\phi}^{P'Q'}\phi_{(CD}\nabla_{B)Q'}\Omega.\end{aligned}$$

With $\phi_{CD}\bar{\phi}_{P'Q'}$ continuous C^1 at $\mathscr{I}$, the above is also continuous at $\mathscr{I}$, so lemma II applies with $\chi_{ABCD} = \Psi_{ABCD}$. Hence ϕ_{ABCD} is continuous at $\mathscr{I}$, and the peeling-off property also holds for the gravitational field, as required.

15. Further discussion

The investigation of zero rest-mass fields described in this paper has been centred on one significant common asymptotic feature they possess, namely the peeling-off property of the p.n.d. associated with each field. While of some considerable interest for its own sake, the real importance of this property perhaps lies in the connexion that these ideas have with some of the more familiar physical quantities associated with radiation. For example, in the case of asymptotically flat space-times, it has become apparent, particularly from the work of Bondi *et al.* (1962),

† For a general space-time the Bianchi identity can be written in terms of $P_{ABC'D'}$ as

$$\nabla^{AR'}\Psi_{ABCD} = -\nabla^{Q'}_{(B}P_{C)DQ'}{}^{R'}$$

Sachs (1962*a*, *b*), Newman & Unti (1962), Goldberg (1963), that the energy-momentum of a system can be measured by examining the $\tilde{r}^{-3}$ part of the gravitational (Weyl curvature) field in the neighbourhood of infinity. (Similarly the total charge is measured by the $\tilde{r}^{-2}$ part of the electromagnetic field.) The rate at which energy-momentum is radiated away by gravitational radiation is measured by the absolute square of the time integral of the radiation field (i.e. of the $\tilde{r}^{-1}$ part) —or alternatively it can be obtained from the $\tilde{r}^{-2}$ part. In the electromagnetic case it is the absolute square of the radiation field itself that enters. In the neutrino case the product of the radiation field with its time-derivative is involved.

All these quantities can be measured by (integrals of) the appropriate expressions on the relevant regions of $\mathscr{I}$ (see Penrose 1963, 1964). In a similar way, one would hope that other quantities such as total angular momentum and multipole structure can be conveniently expressed in terms of integrals over $\mathscr{I}$ (compare Sachs 1962*b*; Goldberg 1963). The relationship between these quantities and the initial data at infinity for the various fields is also of interest. A detailed discussion of some of these matters is left to a later paper.

One point of departure, in this paper, has been the inclusion of asymptotically de Sitter space-times in the discussion of the peeling-off property. However, it is not clear, as yet, to what extent the corresponding discussion of energy-momentum could be carried over to such space-times. Some essential new features are certainly to be expected, in view of the anomalous behaviour of the radiation fields in these cases. Many of the ideas arising out of the conformal technique described here have also a wider application than just to the asymptotically simple space-times. If condition 3 for asymptotically simple space-times is relaxed, then a wide class of cosmological models can be treated. The behaviour of the radiation fields in these models can then be investigated and related to the question of existence of *visual horizons* (Penrose 1964). Furthermore, Norman (1964) has shown that a form of peeling-off can be established for cosmological models in which $\nabla_\mu \Omega = 0$ at $\mathscr{I}$, for fields which satisfy an asymptotic regularity condition. It becomes necessary to replace the affine parameter $\tilde{r}$ by a suitably defined luminosity parameter.

Another related matter that can be treated using this conformal technique is the question of asymptotic groups for curved space-times. The group of conformal self-transformations (appropriately defined) of the hypersurface $\mathscr{I}$ has a covariant significance for the physical space-time $\tilde{\mathscr{M}}$. In this way, it has, for example, been possible to exhibit geometrically (Penrose 1963, 1964) the Bondi–Metzner–Sachs asymptotic symmetry group (Bondi *et al* 1962; Sachs 1962*a*, *b*) which is relevant to asymptotically flat space-times. The inhomogeneous Lorentz group, on the other hand, does not seem to emerge, here, as playing an essential role as an asymptotic symmetry group in general.

The question as to precisely what should be meant by an asymptotically flat space-time has not been fully entered into here. Asymptotic simplicity with $\mathscr{I}$ null would certainly be required. If Einstein's empty-space equations hold near $\mathscr{I}$ then, in addition we have: $C_{\mu\nu\rho\sigma} = 0$ on $\mathscr{I}$; the shear of $\mathscr{I}$ vanishes; $\nabla^{P'}_{(A} \nabla^{Q'}_{B)} \Omega = 0$ on $\mathscr{I}$; in some sense I^0 is a point (rather than, say, a 2-surface). None of these properties need necessarily hold in the absence of some such restriction as $\tilde{R}_{\mu\nu} = 0$ near $\mathscr{I}$.

It is not yet clear how many of these properties ought to be carried over into a definition, if these ideas are to be employed in a precise geometrical statement of asymptotic flatness.

In providing a rigorous covariant definition of both incoming and outgoing gravitational radiation fields, the conformal technique would appear to open the way to a possible 'S-matrix' type of theory for the gravitational field. This, in turn, should be of great value in the construction of a covariant quantum theory of gravitation. However, there are other very relevant and difficult questions, such as of how the gravitational field scatters itself, which have not been touched upon here. (For example, apparently a gravitational wave can effect the scattering of another wave by introducing shear into the wave surface of the second wave.) In addition, of great interest would be an examination of the geometry in the neighbourhood of the point I^0. For this provides the most immediate link between $\mathscr{I}^-$ and $\mathscr{I}^+$. Without such a link, it is difficult to see how the outgoing radiation is to be correlated with the incoming radiation.

The author wishes to thank many of his colleagues for invaluable discussions. In particular, it was partly due to suggestions from E. Schücking that the main idea of this paper came into being. Also the treatment of Hertz potentials given here owes much to discussions with I. Robinson. In addition, suggestions from R. K. Sachs, F. A. E. Pirani and D. Norman have proved very helpful.

The main part of this work was carried out during the author's two-year stay at King's College London and was supported by the Aeronautical Research Laboratory, Office of Air Research, through the European Office of the U.S. Air Force Aerospace Research. Other parts were done at Syracuse University, partly supported by the Aeronautical Research Laboratory while the author was visiting under a N.A.T.O. Fellowship, and at the University of Texas, where it was partly supported by the Air Force Office of Scientific Research contract no. AF-AFOSR-454-63.

References

Arnowitt, R., Deser, S. & Misner, C. 1961 *Phys. Rev.* **121**, 1556.
Bateman, H. 1910 *Proc. Lond. Math. Soc.* **8**, 223.
Bateman, H. 1955 *Electrical and Optical Wave Motion*. New York: Dover.
Bergmann, P. G. 1963 *A.R.L. Tech. Documentary Rep.* 63–56.
Bondi, H. 1960 *Nature, Lond.*, **186**, 535.
Bondi, H., van den Burg, M. G. J. & Metzner, A. W. K. 1962 *Proc. Roy. Soc.* A, **269**, 21.
Buchdahl, H. A. 1959 *Nuovo Cim.* **11**, 496.
Cunningham, E. 1910 *Proc. Lond. Math. Soc.* **8**, 77.
Debever, R. 1959 *C.R. Acad. Sci., Paris*, **249**, 1324, 1744.
Dirac, P. A. M. 1936 *Proc. Roy. Soc.* A, **155**, 447.
Fierz, M. 1940 *Helv. Phys. Acta*, **13**, 45.
Fierz, M. & Pauli, W. 1939 *Proc. Roy. Soc.* A, **173**, 211.
Gårding, L. 1945 *Proc. Camb. Phil. Soc.* **41**, 49.
Goldberg, J. N. 1963 *Phys. Rev.* **131**, 1367.
Goldberg, J. N. & Kerr, R. P. 1964 *J. Math. Phys.* **5**, 172.
Goldberg, J. N. & Sachs, R. K. 1962 *Acta Phys. Polon.* **22**, 13.
Infeld, L. & van der Waerden, B. L. 1933 *S.B. Preuss. Akad.* **9**, 380.
Jordan, P., Ehlers, J. & Sachs, R. 1961 *Akad. Wiss., Mainz*, **1**.
Komar, A. 1962 *Phys. Rev.* **127**, 955.

Kuiper, N. H. 1949 *Ann. of Math.* **50**, 916.
Laporte, O. & Uhlenbeck, G. E. 1931 *Phys. Rev.* **37**, 1380.
McCrae, W. H. 1951 *Proc. Roy. Soc.* A, **206**, 562.
McLennan, J. A. (Jr) 1956 *Nuovo Cim.* **10**, 1360.
Newman, E. T. & Penrose, R. 1962 *J. Math. Phys.* **3**, 566.
Newman, E. T. & Unti, T. W. J. 1962 *J. Math. Phys.* **3**, 891.
Norman, R. D. 1964 Propagation of relativistic gravitational fields. Ph.D. Thesis, London.
Penrose, R. 1959 Report to the Royaumont Conference.
Penrose, R. 1960 *Ann. of Phys.* **10**, 171.
Penrose, R. 1962 Report to the Warsaw Conference.
Penrose, R. 1963 *Phys. Rev. Lett.* **10**, 66.
Penrose, R. 1964 article in *Relativity, groups and topology: the* 1963 *Les Houches lectures*. New York: Gordon and Breach.
Pirani, F. A. E. 1957 *Phys. Rev.* **105**, 1089
Pirani, F. A. E. 1959 *Proc. Roy. Soc.* A, **251**, 519.
Robinson, I. & Trautman, A. 1962 *Proc. Roy. Soc.* A, **265**, 463.
Rudberg, H. 1957 The compactification of a Lorentz space and some remarks on the foundation of the theory of conformal relativity. Thesis, Uppsala; 1958 *Phys. Abstr.* no. 30, vol. 61.
Ruse, H. S. 1937 *Proc. Roy. Soc. Edinb.* **57**, 97.
Ruse, H. S. 1944 *J. Lond. Math. Soc.* **19**, 168.
Sachs, R. 1961 *Proc. Roy. Soc.* A, **264**, 309.
Sachs, R. 1962*a* *Proc. Roy. Soc.* A, **270**, 103.
Sachs, R. 1962*b* *Phys. Rev.* **128**, 2851.
Sachs, R. & Bergmann, P. G. 1958 *Phys. Rev.* **112**, 674.
Schouten, J. 1954 *Ricci Calculus*. Berlin: Julius Springer.
Synge, J. L. *Relativity: the special theory*. Amsterdam: North-Holland.
Szekeres, P. 1963 *Proc. Roy. Soc.* A, **274**, 206.
Trautman, A. 1958 *Bull. Acad. Polon. Sci.* **6**, 407.
van der Waerden, B. L. 1929 *Nachr. Ges. Wiss. Gottingen*, **100**.
Wigner, E. P. & Bargmann, V. 1948 *Proc. Nat. Acad. Sci., Wash.*, **34**, 211.
Witten, L. 1959 *Phys. Rev.* **115**, 206.

Appendix

Let the space-time $\tilde{\mathscr{M}}$ and corresponding bounded manifold $\mathscr{M}$ satisfy the conditions 1, 2, 3, 4 for asymptotic simplicity laid down in §9. Suppose, further, that $\dot{\mathscr{M}} = \mathscr{I} = \mathscr{I}^+ \cup \mathscr{I}^-$ is everywhere null. It is required to show (in outline) that $\mathscr{I}^+ \cong \mathscr{I}^- \cong S^2 \times E^1$.

Let Q be any point in the interior of $\mathscr{M}$. Let $\mathscr{C}$ be the complete future null cone of Q, that is, $\mathscr{C}$ is generated by the totality of null geodesics with past endpoint Q and future endpoints on $\mathscr{I}^+$. Let $\mathscr{Q}$ denote the set of points of $\mathscr{M}$ to the future of Q, so that $P \in \mathscr{Q}$ if and only if there exists a time-like curve which is future oriented from Q to P (and $P \neq Q$). Let $\mathscr{B}$ denote the closure, in $\mathscr{M}$, of the boundary of $\mathscr{Q}$ in the interior of $\mathscr{M}$, i.e. $\mathscr{B} = \dot{\mathscr{Q}} - \mathscr{Q} \cap \mathscr{I}$. From local considerations it is easy to see that the hypersurface $\mathscr{B}$ is everywhere null, at some places consisting, perhaps, of the *past portions* of the intersections of null hypersurfaces—and near Q it coincides with $\mathscr{C}$ (since closed time-like curves are excluded). Thus, if $P \in \mathscr{B}$ and $P \neq Q$, then there will be one or more null geodesic segments on $\mathscr{B}$ with future endpoint P. Furthermore, following such a null geodesic back into the past on $\mathscr{B}$, we must eventually arrive back at Q—or else, conceivably, at some point B on $\mathscr{I}^-$. However, this second possibility cannot arise, since if $\mathscr{B}$ meets $\mathscr{I}^-$, any curve into the past, destined for Q, and starting from a point just to the future of B, would have either to terminate

on $\mathscr{I}^-$ or cross $\mathscr{B}$. It therefore follows that the hypersurface $\mathscr{B}$—the future *causal* null cone of Q—is *part* (or perhaps all) of the complete future cone $\mathscr{C}$. Clearly $\mathscr{C} - \mathscr{B} \subset \mathscr{Q}$.

In fact, $\mathscr{B}$ is generated by the null geodesic segments with past endpoint Q, which are extended into the future along $\mathscr{C}$ until either $\mathscr{I}^+$ or a singularity of $\mathscr{C}$ is reached. Thus, $\mathscr{B}$ consists of all points of $\mathscr{C}$ which can be connected to Q by a curve on $\mathscr{C}$ not crossing any other singularities of $\mathscr{C}$. The singularities of $\mathscr{C}$ will be self-intersection regions and cuspidal regions (both generally two-dimensional), but $\mathscr{C}$ will be a *compact* set (with boundary) and will possess a finite number of non-singular parts. $\mathscr{B}$, also, will therefore be a compact set (with one connected non-singular part) and $\dot{\mathscr{B}} = \mathscr{B} \cap \mathscr{I}^+$.

Until now, the fact that $\mathscr{I}$ is null has not been used in an essential way. It will next be shown that any generator h of $\mathscr{I}^+$ which meets $\mathscr{C}$ must also meet $\mathscr{B}$. For, suppose h does not meet $\mathscr{B}$ but it meets a generator s of $\mathscr{C}$ at R. Then $R \in \mathscr{Q}$. Let S be a variable point of s and choose a null geodesic $g(S)$ through each point S in a continuous way with $g(R) = h$. (For example, we could define each $g(S)$ by requiring that its tangent at S be parallel, along s, to that of h at R.) Now, clearly there are points near $g(Q) \cap \mathscr{I}^-$ which are not in $\mathscr{Q}$ (since closed time-like curves are excluded). Also we have seen that $\mathscr{B}$ does not meet $\mathscr{I}^-$. Hence, points of $g(S)$ sufficiently near $\mathscr{I}^-$ are not in $\mathscr{Q}$, for each S on s ($S \neq R$). But when S is sufficiently close to R say between S_1 and R, then $S \in \mathscr{Q}$. Thus $g(S)$ meets $\mathscr{B}$ in a point $B(S)$ whenever S lies between S_1 and R. The correspondence $S \to B(S)$ maps the open segment $S_1 R$ of s continuously into the compact set $\mathscr{B}$. Therefore there is a limit point T on $\mathscr{B}$ with the property that $B(S)$ becomes arbitrarily close to T for some points S sufficiently close to R. Furthermore, since the null directions at a point also form a compact set, it follows that there is a limiting null geodesic t through T with the property that $g(S)$ lies arbitrarily close to t for some points S sufficiently close to R. But $t \neq h$, since we have assumed that $\mathscr{B}$ does not meet h. This is a contradiction since $g(S)$ approaches h as S approaches R.

Thus each generator of $\mathscr{I}^+$ which meets $\mathscr{C}$ must also meet $\mathscr{B}$. Moreover, it meets $\mathscr{B}$ in exactly one point and has the topology E^1 (rather than S^1, since after entering $\mathscr{Q}$ it cannot leave $\mathscr{Q}$ again while continuing in a future direction). The generators of $\mathscr{I}^+$ meeting $\mathscr{C}$ therefore constitute a manifold which has the topology $\dot{\mathscr{B}} \times E^1$. But since $\dot{\mathscr{B}}$ is necessarily a compact manifold without boundary, any other part $\mathscr{I}^*$ of $\mathscr{I}^+$ would have to be disconnected from the $\dot{\mathscr{B}} \times E^1$ part. It follows that any point Q^* whose complete future null cone meets $\mathscr{I}^*$ must be disconnected from Q. This is impossible since $\mathscr{M}$ is connected. Hence $\mathscr{I}^+ \cong \dot{\mathscr{B}} \times E^1$.

It remains to show that $\dot{\mathscr{B}} \cong S^2$. Now the set of future null directions at Q has the topology S^2. Each such null direction determines a point on $\mathscr{I}^+$ namely the intersection of the corresponding generator of $\mathscr{C}$ with $\mathscr{I}^+$. Thus we have a continuous map $S^2 \to \mathscr{I}^+$, the image under the map being $\dot{\mathscr{C}} = \mathscr{I}^+ \cap \mathscr{C}$. Since $\mathscr{C}$ will generally possess cuspidal 2-surfaces and self-intersection regions, it follows that the 2-surface $\dot{\mathscr{C}}$ will generally also be self-intersecting and possess cuspidal edges. A positive or negative sign can be attached to each non-singular part of $\dot{\mathscr{C}}$ according as areas are described (relative to the null direction in $\mathscr{C}$) in the same or opposite sense as the

corresponding areas on the smooth S^2 which represents the future null directions at Q. Thus, the sign on $\dot{\mathscr{C}}$ changes whenever a cuspidal edge is crossed: the cuspidal edges are curves along which the surface reverses its sense.

There is a certain integer invariant $I(\mathscr{S})$ which can be associated with any closed consistently signed 2-surface $\mathscr{S}$ which, like $\dot{\mathscr{C}}$ above, may possess cuspidal edges and self-intersection curves. To find $I(\mathscr{S})$ we simply build up $\mathscr{S}$ out of *smooth* S^2's some of which are positively signed and others negatively signed. (This can in general be done, but exceptionally $\mathscr{S}$ may possess nodes. *Approximate* such an $\mathscr{S}$ by smooth S^2's.) Where a portion of a positively signed S^2 coincides with a portion of a negatively signed S^2 the two portions are considered to cancel out. $I(\mathscr{S})$ is then the net number (positive minus negative) of signed S^2's involved. To see that $I(\mathscr{S})$ is indeed an invariant, independent of the way in which $\mathscr{S}$ has been built up from the S^2's, consider the integral with respect to surface area—negative portions counting negatively—of the Gaussian curvature of the non-singular parts of $\mathscr{S}$. The contribution from each smooth S^2 is $\pm 4\pi$ (according to its sign) by Gauss's theorem. Hence the total integral for $\mathscr{S}$ is $4\pi I(\mathscr{S})$.

If we choose a parameter r on each generator of $\mathscr{C}$ which takes the value zero at Q and unity at $\mathscr{I}^+$ and which varies smoothly from generator to generator, the points on $\mathscr{C}$ corresponding to a fixed value of r will describe a 2-surface $\mathscr{S}_r$ of the kind mentioned above, and $\mathscr{S}_1 = \dot{\mathscr{C}}$. The invariant $I(\mathscr{S}_r)$, being a continuously varying integer, must remain constant for all r. Hence $I(\dot{\mathscr{C}}) = 1$, since clearly $I(\mathscr{S}_\epsilon) = 1$ for small $\epsilon > 0$. If the 2-surface $\mathscr{S}$ is projected down along generators of $\mathscr{I}^+$, we get a map $\dot{\mathscr{C}} \to \dot{\mathscr{B}}$. In fact, we can clearly accomplish this projection in a continuous way along $\mathscr{I}^+$, that is, for each $p (0 \leqslant p < 1)$ we can define a closed signed 2-surface $\mathscr{T}_p$ (with cuspidal edges, etc.) on $\mathscr{I}^+$, which varies continuously with p from $\dot{\mathscr{C}} = \mathscr{T}_0$ until a limiting position on $\dot{\mathscr{B}}$ is reached. Hence $1 = I(\dot{\mathscr{C}}) = I(\mathscr{T}_p) = nI(\dot{\mathscr{B}})$ where the *integer* n is the degree of the map $\dot{\mathscr{C}} \to \dot{\mathscr{B}}$ (number of points of $\dot{\mathscr{C}}$ which map to a point of $\dot{\mathscr{B}}$ counted algebraically). It follows that $n = \pm 1$. The existence of a continuous map $S^2 \to \dot{\mathscr{B}}$ of degree ± 1 is thus established and it follows that $\dot{\mathscr{B}} \cong S^2$ as required. The proof that $\mathscr{I}^- \cong S^2 \times E^1$ is similar.

23
10 EXACT GRAVITATIONALLY-CONSERVED QUANTITIES

Phys. Rev. Letters **15(6)** (1965), 231–233.
With E. T. Newman

The actual significance of the curious conservation laws ('NP-constants') that we introduced in this paper has, in my opinion, never properly come to light. These laws have been widely misunderstood—even by some very distinguished people; see for example Misner, C.W., Thorne, K.S., and Wheeler, J.A. (1973) *Gravitation* (Freeman, San Francisco), pp. 870–871, ex.32.11 and Bondi, H. (1967) *Assumption and Myth in Physical Theory* (Cambridge University Press, Cambridge). For further developments and more explanation, see Chapters 27, 30, 35, and 41; also R. Penrose and W. Rindler (1986) *Spinors and Space-Time*, Volume 2: *Spinor and Twistor Methods in Space-Time Geometry* (Cambridge University Press, Cambridge), pp. 427–9.

PHYSICAL REVIEW
LETTERS

VOLUME 15 | 9 AUGUST 1965 | NUMBER 6

10 EXACT GRAVITATIONALLY-CONSERVED QUANTITIES*

E. T. Newman†
University of Pittsburgh, Pittsburgh, Pennsylvania

and

R. Penrose
Birkbeck College, University of London, London, England
(Received 28 June 1965)

In ordinary electromagnetic theory, total charge remains exactly constant for any system surrounded by source-free space. This is a direct consequence of Maxwell's equations for the free-space region, whereas dipole and higher moments can be carried away by electromagnetic radiation. In gravitational theory, analogs of this conserved charge would, at first sight, appear to be mass, momentum, and angular momentum, as they are indeed exactly conserved in the linearized Einstein theory, for regions surrounded by empty space.[1,2] However, in the full nonlinear theory mass and momentum can be changed (i.e., "carried away") by gravitational radiation.[3-5] It has therefore been commonly believed that there are no (nontrivial) exactly conserved (covariant) gravitational quantities in general relativity. But it turns out that, quite unexpectedly, there is a set of 10 geometrical quantities, defined for asymptotically flat space-times, which have a quadrupole structure, and whose values cannot be altered in any way by gravitational radiation.

For a general asymptotically flat space-time[3-6] the quantities can be expressed as integrals at infinity on any null hypersurface $\mathfrak{N}$ which diverges to infinity. Using the notation of Newman-Unti,[5,7] these expressions are

$$Q_m = \int\int \Psi_0{}^1 {}_2Y_{2,m} \sin\theta d\varphi d\theta,$$
$$m = -2, -1, 0, 1, 2. \qquad (1)$$

Here $\Psi_0{}^1$ is defined by $-C_{\mu\nu\rho\sigma}l^\mu m^\nu l^\rho m^\sigma = \Psi_0 = \Psi_0{}^0 r^{-5} + \Psi_0{}^1 r^{-6} + O(r^{-7})$, where r is an affine (or luminosity) parameter, with $(l^\mu, n^\mu, m^\mu, \bar{m}^\mu)$ as standard null tetrad, on $\mathfrak{N}$, such that $r^2(d\theta^2 + \sin^2\theta d\varphi^2)$ defines, for large r, the metric of the "sphere at infinity" on $\mathfrak{N}$ ($\Psi_0{}^0, \Psi_0{}^1$ are independent of r). The quantities ${}_2Y_{2,m}$ are effectively "tensor" spherical harmonics[2] defined in terms of the ordinary harmonics $Y_{2,m}$ by

$${}_2Y_{2,m} = \left(\sin\theta\frac{\partial}{\partial\theta} + i\frac{\partial}{\partial\varphi}\right)\left(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta}\frac{\partial}{\partial\varphi}\right)Y_{2,m}.$$

The real and imaginary parts of the five complex quantities Q_m are the required conserved quantities. Together they constitute a representation $D(2,0)$ of the homogeneous Lorentz group [of conformal motions of the (θ,φ) sphere].[4] They are conserved in the sense that if space-time is empty near infinity and if $\mathfrak{N}'$ is any other null hypersurface diverging to infinity, then the new values of Q_m will be the same as the old, the θ, φ for $\mathfrak{N}'$ being carried over from

those for $\mathfrak{N}$ in the standard way, i.e., by the "supertranslation" of $\mathfrak{N}$ into $\mathfrak{N}'$.[3,4] Thus, the Q_m also form a $D(2,0)$ representation[8] of the full Bondi-Metzner-Sachs group.[3,4,9] This assumes that $\mathfrak{N}$ and $\mathfrak{N}'$ both open out into the future, so that outgoing gravitational radiation might escape between them. We could also define 10 additional conserved quantities by using (1) on null hypersurfaces which open out into the past. These would, in general, be different from the above Q_m, but the fact that they are conserved would seem to be of lesser interest, for retarded fields.[6]

It may be asked what is the interpretation of (1) in the linearized theory. Again the Q_m would be conserved, but in this case we must interpret (1) as representing the presence of incoming radiation (assuming $\mathfrak{N}$ opens into the future). It can be shown that (1) in fact vanishes for all retarded fields in the linearized theory. It would be natural, therefore, to attempt to interpret (1) in terms of incoming radiation in the full theory also. However, it is here that the full theory presents us with an essentially new situation. As it stands, (1) does not represent incoming radiation alone. This is brought out most clearly by an examination of the static and stationary vacuum fields. In such cases incoming radiation is absent (according to any reasonable definition!), but the Q_m can be expressed as certain combinations of mass, dipole, and quadrupole moments. For, if in (1) we replace Ψ_0^1 by

$$-5\Psi_0^0\Psi_2^0+(10/3)(\Psi_1^0)^2, \qquad (2)$$

the field being stationary, and we choose $\mathfrak{N}$ (as we now may) so that the leading term σ^0 in the shear vanishes, then it turns out that the same Q_m are still obtained. Here Ψ_2^0 is real, constant, and represents the total mass; Ψ_1^0, Ψ_0^0 are complex functions of θ and φ describing, respectively, dipole and quadrupole structure of the source.[5,2] [Expression (2) has a quadrupole angular structure and is now equivalent to the Q_m.] In the case of the Weyl-Levi-Civita solutions only Q_0 survives and is real, but it does not vanish in general, i.e., unless the particular combination of moments given by (2) vanishes.

This leads to a remarkable state of affairs. Suppose an initially asymmetric body [with (2) nonvanishing] becomes spherically symmetric at a later time. In the meantime it must radiate gravitationally and the quantity (2) may change, but if the system is ever to become stationary again, the combination of moments (2) would have to return to its original value! It appears that we must interpret this in terms of back-scattering of the gravitational waves. Thus, during the radiative period, (1) represents not only a multipole structure of the source, but, in addition, a part apparently corresponding to incoming (i.e., scattered) radiation which would destroy the stationary nature of the field of the body—unless and until the original values of (2) are regained! That incoming radiation can contribute to the Q_m is clear from the results of the linear theory.

It should be mentioned here that while the Q_m determine a selection rule for purely gravitational transitions between stationary states, in the presence of electromagnetism (or neutrinos) the Q_m are no longer conserved. (It is conceivable that this has relevance to gravitational collapse, since a transition which cannot be accomplished by gravitational radiation alone might take place electromagnetically!) However, when electromagnetism is present, six new exactly conserved quantities emerge, namely the analogs of (1) for the Maxwell field. They have a dipole structure and depend on the r^{-4} part of the Maxwell field; they apparently do not necessarily vanish for stationary fields and they are conserved in asymptotically flat space-times if the Einstein-Maxwell equations hold near infinity. (Also there are four or eight conserved quantities for neutrino fields.)

Full details of these results, with proofs, will appear elsewhere.

---

*Work supported in part by the Aerospace Research Laboratory, through the European Office of the U. S. Air Force (and, in part, by Contract No. AF-AFOSR-454-63).

†On temporary leave at King's College, University of London, London, England.

[1]R. K. Sachs and P. G. Bergmann, Phys. Rev. 112, 674 (1958).

[2]A. I. Janis and E. T. Newman, to be published.

[3]H. Bondi, Nature 186, 535 (1960); H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, Proc. Roy. Soc. (London) A269, 21 (1962).

[4]R. K. Sachs, Proc. Roy. Soc. (London) A270, 103 (1962).

[5]E. T. Newman and T. W. J. Unti, J. Math. Phys. 3, 891 (1962).

[6]For an alternative, though equivalent, approach to asymptotic flatness in which a discussion of both in-

coming and outgoing fields is made, see R. Penrose, Proc. Roy. Soc. (London) A284, 159 (1965).

[7]E. T. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962); 4, 998 (1963).

[8]E. T. Newman, Nature 206, 811 (1965).

[9]R. K. Sachs, Phys. Rev. 128, 2851 (1962).

24
ENERGY CONSERVATION AS THE BASIS OF RELATIVISTIC MECHANICS II

Am. J. Phys. **33(12)** (1965), 995–997.
With W. Rindler and J. Ehlers

This is an improved version of Chapter 19.

Energy Conservation as the Basis of Relativistic Mechanics. II*

J. EHLERS, W. RINDLER
Southwest Center for Advanced Studies, Dallas, Texas 75230
AND
R. PENROSE
Birkbeck College, London, England
(Received 10 May 1965)

From the relativity principle and the conservation of energy in particle collisions we deduce the form of the energy function, and the conservation of inertial mass and three-momentum. We show that the arguments are parallel under Einsteinian and Galilean kinematics.

A PAPER with the above title was recently published in this Journal by two of the authors.[1] In the present paper a modified argument is given which represents a significant improvement in brevity and precision. As the argument now stands, it would appear to be the simplest and pedagogically most satisfactory approach to (Einsteinian) relativistic mechanics suggested so far. For this reason, rather than merely communicate the changes from the previous paper, we thought it best to make the present account self-contained, especially since the overlap is slight. At the same time we have made brief mention of the parallel development of Newtonian mechanics. For a historical introduction and bibliography we still refer the reader to the earlier paper.

Our notation is the following. (It differs slightly from that of Ref. 1.) The magnitude of the 3-velocity $\mathbf{u}_F$ of a particle relative to an inertial frame F is denoted by u_F (the speed). The 4-velocity corresponding to any $\mathbf{u}$ is denoted by $\mathbf{U}$, and $\mathbf{U}\cdot\mathbf{V}$ means the Minkowskian scalar product $c^2U^4V^4-U^1V^1-U^2V^2-U^3V^3$ of two 4-vectors $\mathbf{U}=(U^1,U^2,U^3,U^4)$ and $\mathbf{V}=(V^1,V^2,V^3,V^4)$. The internal state of a particle, comprising all such properties as material constitution, spin, electric charge, temperature, etc., which can be measured in a rest frame of the particle, is indicated by a collective symbol $\mathcal{S}$. In order to divest $\mathcal{S}$ of its directional properties, we define *scalar* states S as equivalence classes of states $\mathcal{S}$, such that distinct $\mathcal{S}$ belong to the same S if they have the same description in suitably oriented rest frames.—Greek *superscripts* distinguish the states, velocities, etc., of different particles.

The mechanics to be discussed is that of particle collisions. We assume that the particles are "free," i.e., that their motions are independent of each other except during collisions, and that before and after collision each particle moves uniformly in accordance with the law of inertia (which is already incorporated into special relativity in the definition of inertial frames).

We make the following assumptions.

Assumption (I): There exists a real-valued universal energy function $E(u,S)$, where u ranges over-all speeds and S over-all scalar states, such that in any collision the energy sum relative to any inertial frame F is conserved:

$$\sum_{\alpha=1}^{n} E(u_F{}^\alpha,S^\alpha)=\sum_{\alpha=n+1}^{n+m} E(u_F{}^\alpha,S^\alpha) \qquad (1)$$

(the particles numbered from 1 to n go into the collision, and those numbered from $n+1$ to $n+m$ come out of it). For any fixed S, $E(u,S)$ is a continuous nonconstant function of u.

The last property distinguishes an energy function from an "internal charge," i.e., a function of S alone whose sum is conserved in any collision. Since such an internal charge can be added to E without invalidating (1), it is clear that (1) does not define E uniquely.

A change of scalar state which occurs in a free particle without external influence can be considered as a one-into-one "collision." By the law of inertia such a change does not affect u, and by (1) it does not affect E. Hence, for the pur-

* This work was supported in part by NASA Grant No. NsG269-62 and in part by the U. S. Air Force, Office of Aerospace Research: AFOSR under Grant No. AF-AFOSR-903-65 and ARL under Contract No. AF 33(615)-1029.

[1] R. Penrose and W. Rindler, Am. J. Phys. **33**, 55 (1965).

poses of our analysis, such variable states can be considered as single states.

Assumption (II): The sum of the energies of two identical particles approaching each other with equal speeds u along any straight line in an inertial frame F is independent of this line, even relative to any other frame F'.

This is the analog, for the simplest of all 2-particle systems, of the direction-independence assumed for the energy of a single particle. Although this assumption cannot be deduced from Assumption (I), it is in accordance with the general energy principle. For we may suppose that the particles transfer their energy to a symmetric system at rest in F whose gain in energy relative to F must be the same in all cases; thus the gain relative to F' must also be the same.

Assumptions (I) and (II) can be used in conjunction with either Galilean or Einsteinian kinematics. We first consider the Einsteinian case.

For convenience, we introduce instead of u a new variable, namely the Lorentz factor

$$\gamma=\gamma(u)=(1-u^2/c^2)^{-\frac{1}{2}}; \tag{2}$$

and we temporarily write $E(u,S)\equiv E[\gamma]$ when only one scalar state is under discussion. We recall that the 4-velocity associated with a 3-velocity $\mathbf{u}$ is given by $\mathbf{U}=\gamma(u)(\mathbf{u},c)$; hence the 4-velocity of an inertial observer relative to his own rest-frame is $\mathbf{V}=(\mathbf{0},c)$. Thus

$$\gamma=\mathbf{V}\cdot\mathbf{U}/c^2, \tag{3}$$

i.e., the Lorentz factor of a particle relative to any inertial observer is given by the scalar product of the two corresponding 4-velocities divided by c^2. And this product, being invariant, can be evaluated in any inertial frame.

Now consider two particle systems as in Assumption (II), along the x and y axes, respectively. The components of the four relevant 4-velocities with respect to F are

$$\gamma(\pm u,\,0,\,0,\,c);\quad \gamma(0,\,\pm u,\,0,\,c);\quad [\gamma=\gamma(u)]. \tag{4}$$

If F' moves with speed u' along the x axis of F, its 4-velocity components with respect to F are

$$\gamma'(u',0,0,c),\quad [\gamma'=\gamma(u')]. \tag{5}$$

Hence, the various Lorentz factors of the particles relative to F' are given by the scalar products (divided by c^2) of (5) with the 4-velocities in (4):

$$\gamma_1=\gamma\gamma'(1+uu'/c^2),\quad \gamma_2=\gamma\gamma'(1-uu'/c^2),$$
$$\gamma_3=\gamma_4=\gamma\gamma'=\tfrac{1}{2}(\gamma_1+\gamma_2). \tag{6}$$

Equating the energies of the two systems in F' in accordance with Assumption (II), we obtain the functional equation

$$E[\gamma_1]+E[\gamma_2]=2E[\tfrac{1}{2}(\gamma_1+\gamma_2)], \tag{7}$$

valid for all values $1\leq\gamma_1,\ \gamma_2<\infty$ since u and u' can be varied independently. In words, this equation reads: "the average of E at γ_1 and γ_2 equals E at the average of γ_1 and γ_2." In the usual way (continued halving and appeal to the continuity of E) it follows that E must be a linear function of γ

$$E=c^2m\gamma+q, \tag{8}$$

where m and q are functions of S (the factor c^2 being introduced for later convenience) and, because of the rider in Assumption (I),

$$m(S)\neq 0 \quad \text{for any } S. \ ^{(2)} \tag{9}$$

Now consider an arbitrary collision and let the 4-velocities of the particles involved be $\mathbf{U}^\alpha$. If $\mathbf{V}$ is the 4-velocity of an arbitrary inertial frame F (i.e., its "time axis") then according to (8) and (3) the energy sum relative to F can be written as

$$\mathbf{V}\cdot\sum_\alpha m^\alpha\mathbf{U}^\alpha+\sum_\alpha q^\alpha. \tag{10}$$

Since this expression is conserved for each 4-velocity $\mathbf{V}$ it follows that $\sum m^\alpha\mathbf{U}^\alpha$ and $\sum q^\alpha$ are *separately* conserved. [Take, for example, $\mathbf{V}$ as in (5) and allow u' to vary. Then the conservation of (10) implies that of $\sum mU^1$, $\sum mU^4$, and $\sum q$; etc.] Note that q has therefore the character of an internal charge and mU^4 that of an energy function. We infer

Theorem (I): Under Einsteinian kinematics, Assumptions (I) and (II) are equivalent to the postulate that there exists a nonvanishing func-

[2] It would be interesting to know what further assumptions are sufficient within a classical collision theory to imply that m must have the same sign for all particles.

tion m of S such that the 4-vector sum

$$\sum_\alpha m^\alpha \mathbf{U}^\alpha \tag{11}$$

is conserved in all collisions.

We can now define the proper mass of a particle as the scalar m, and its 4-momentum as the 4-vector $m\mathbf{U}$, provided that m is unique (up to a universal factor corresponding to a change of unit). The conservation of 4-momentum, the central law of relativistic collision mechanics, would then be established.

Without attempting to give the most general uniqueness argument for m, we only state one simple assumption which is sufficient: Any two particles P and P' can be made either to collide and stick together upon impact; or to collide so that two particles again emerge and three of the four 4-velocities involved are linearly independent; or a finite number of particles $P_1, P_2, \cdots, P_n$ can be found such that each of the pairs P,P_1; P_1,P_2; $\cdots$; P_n,P' can undergo such a collision. For in any such collision the (observable) 4-velocities determine the *ratios* of the m uniquely. If there exist particles which can participate only in more complicated (e.g., 5-particle) collisions, the above assumption and argument would have to be somewhat extended.

The 3-momentum of a particle relative to a frame F can now be defined in the usual way as $m\gamma\mathbf{u}_F$ and the inertial mass as $m\gamma$, and both of these are conserved in any frame (the first being the spatial component and the second $1/c$ times the temporal component of the 4-momentum.)

Furthermore, $E=c^2m\gamma$ is seen to be *an* energy function satisfying Assumptions (I) and (II). This is evidently the most natural choice, since the addition of a charge-like quantity q as in (8) produces a hybrid entity, $c^2m\gamma$ transforming as the last component of a 4-vector ($cm\mathbf{U}$) and q as a scalar. On the other hand, it appears to be outside the scope of special relativity to prove that $c^2m\gamma$ is the *available* (as distinct from *conventional*) energy of a mass m. If, for example, the elementary particles were indestructible, the available energy of a macroscopic particle would be $c^2m\gamma+q$, where q is $-c^2$ times the sum of the rest masses of the constituent elementary particles. However, in general relativity, a kind of proof can be given by lowering—in imagination—a particle into the "Schwarzschild radius" of a mass point.

The basis of special relativistic mechanics is now established, and we need not discuss it further.

We finally formulate briefly the corresponding argument if Galilean kinematics is adopted. It is then convenient to use u^2 as a new variable instead of u; in place of (6) we obtain

$$u_1^2=(u-u')^2, \quad u_2^2=(u+u')^2,$$
$$u_3^2=u_4^2=u^2+u'^2=(u_1^2+u_2^2)/2, \tag{6'}$$

and in place of (7),

$$E[u_1^2]+E[u_2^2]=2E[\tfrac{1}{2}(u_1^2+u_2^2)]. \tag{7'}$$

Except for notation, this is identical with (7). Hence

$$E=\tfrac{1}{2}mu^2+E_0, \tag{8'}$$

where $m(\neq 0)$ and E_0 are functions of S, and the factor $\frac{1}{2}$ is included so that (8) and (8') can coincide in the limit of small velocities. Instead of (10), the Galilean relativity principle leads to the conservation of

$$\sum \tfrac{1}{2}m^\alpha(\mathbf{u}^\alpha+\mathbf{u}')^2+\sum E_0{}^\alpha \equiv \sum(\tfrac{1}{2}m^\alpha u^{\alpha 2}+E_0) + \mathbf{u}'\cdot\sum m^\alpha\mathbf{u}^\alpha+\tfrac{1}{2}u'^2\sum m^\alpha \tag{10'}$$

for each $\mathbf{u}'$ and, consequently, to the conservation of the quantities $\sum m^\alpha\mathbf{u}^\alpha$ and $\sum m^\alpha$, *separately*. A theorem analogous to Theorem (I) can therefore be formulated. It is also obvious that the analog of the argument for the uniqueness of m holds here.

The development (6')–(10') is essentially a "limit" of the special relativistic case, for small velocities. This becomes clear when it is remembered that $\gamma(u)=1+\frac{1}{2}u^2/c^2+O(1/c^4)$.

Thus in the Galilean as well as in the Einsteinian case, an energy conservation law in conjunction with the relativity principle implies a 3-momentum conservation law and the law of conservation of inertial mass. In the Einsteinian case, however, the energy function possesses a natural zero point whereas, since the two components of E in (8') are not separately conserved, no such zero point exists in the Galilean case. In this respect, which has a deep physical significance, Einstein's theory is more determinate than its predecessor.

25
GENERAL-RELATIVISTIC ENERGY FLUX AND ELEMENTARY OPTICS

1966 *Perspectives in Geometry and Relativity (essays in Honor of Václav Hlavatý).* Ed. Banesh Hoffmann pp. 259–274. Indiana University Press, Bloomington

In this paper I explored the connection between the Bondi formula for mass loss through gravitational radiation and some basic considerations of ordinary optics (related to the gravitational lensing effects that have now become important in astronomy). See Chapter 140 for a related treatment of angular momentum loss.

 General-relativistic energy flux and elementary optics. 1966 *Perspectives in Geometry and Relativity (essays in Honor of Václav Hlavatý). Ed. Banesh Hoffmann pp. 259–274*

[27]

GENERAL-RELATIVISTIC ENERGY FLUX AND ELEMENTARY OPTICS

Roger Penrose

From:
Perspectives in Geometry and Relativity
[Essays in Honor of Václav Hlavatý]

Copyright © 1966 by Indiana University Press

General-relativistic energy flux and elementary optics. 1966 *Perspectives in Geometry and Relativity (essays in Honor of Václav Hlavatý). Ed. Banesh Hoffmann pp. 259–274*

27. General-Relativistic Energy Flux and Elementary Optics*

ROGER PENROSE
Birkbeck College, London University, London, England

Abstract. It is suggested that the resultant focusing power of space-time curvature along a null ray is a good measure of the total energy flux (matter plus gravitation) across the ray. The effect of Weyl curvature is compared with that of purely astigmatic lenses, and nonlocal behavior is exhibited which reflects the essential non-locality of gravitational energy. A precise measure of the focusing power of a system is suggested, and this is compared with the asymptotic expressions for the energy carried away by zero-mass radiation.

1. Nonlocality of Energy

In view of Professor Hlavatý's many contributions to the understanding of geometrical matters in general-relativistic field theories, it would seem appropriate here to present a discussion with a strong geometrical flavor, but which, in addition, could possibly have important physical implications in the general theory of relativity. I propose to indicate, in fact, that certain elementary considerations of ordinary lens optics may be of value for the further understanding of that somewhat elusive concept—gravitational energy. The discussion given here will be partly a heuristic one designed to bring out certain analogies and to indicate the general applicability of the main idea. Whether or not this idea can be made completely rigorous and physically acceptable will depend on future developments. At the moment, what can be said definitely is that it is correct in a certain limiting sense.

Roughly speaking, the idea is that, given any null ray (that is, a null geodesic) in space-time, the energy flux (gravitational plus matter) across the ray can be measured by the resultant focusing of an idealized light beam following the null ray, due to the space-time curvature.

When we come to examine this in a little more detail shortly, it will be necessary to consider the separate effects of the two main types of curvature. The Riemann tensor (sign chosen so that $k_{\lambda;\mu;\nu} = R_{\lambda\alpha\mu\nu}K^{\alpha}$; space-time signature $(+---)$) can be decomposed as follows:

$$R_{\mu\nu}{}^{\rho\sigma} = C_{\mu\nu}{}^{\rho\sigma} + \delta^{[\rho}_{[\mu}\, S^{\sigma]}_{\nu]} \qquad (1)$$

---

* This research was carried out partly in 1960-1961 while the author was a recipient of a NATO Science Fellowship at Princeton University, and also at Cornell University, where support was given by the National Science Foundation.

(square brackets denoting the skew part), where

$$S_{\mu\nu} = 2R_{\mu\nu} - \tfrac{1}{3}Rg_{\mu\nu},\ R_{\mu\nu} = R_{\mu\alpha\nu}{}^{\alpha},\ \text{and}\ R = R_{\alpha}{}^{\alpha}, \tag{2}$$

so that $C_{\mu\nu\rho\sigma}$ is Weyl's conformal tensor satisfying $C_{\mu\nu\rho\sigma} = C_{[\rho\sigma][\mu\nu]}$, $C_{\mu\nu\rho}{}^{\nu} = 0$, and $C_{\mu[\nu\rho\sigma]} = 0$. The Weyl tensor may be thought of as that part of the curvature representing the gravitational degrees of freedom, the part due to Ricci tensor, i.e., to "matter," having been subtracted out. The local matter stress-energy density is given, according to Einstein, by

$$T_{\mu\nu} = -\kappa(R_{\mu\nu} - \tfrac{1}{2}Rg_{\mu\nu} - \lambda g_{\mu\nu}), \tag{3}$$

where κ and λ are constant with $\kappa > 0$ (and we may choose $\lambda = 0$). Units will be chosen here so that $\kappa = \frac{1}{2}$. The information contained in $S_{\mu\nu}$ is thus the same as that contained in $T_{\mu\nu}$. There is no place for a covariant local energy density of gravitational energy appearing here. By definition, $T_{\mu\nu}$ describes all the local energy, so any energy due to $C_{\mu\nu\rho\sigma}$ must be nonlocal in character. The existence of nonlocal energy is also manifest in the fact that the local conservation law,

$$T_{\mu\nu}{}^{;\,\nu} = 0, \tag{4}$$

is a covariant curved-space equation. One cannot just integrate $T_{\mu\nu}$ over a 3-space to obtain a conserved total energy-momentum. The total energy-momentum of a system must, therefore, involve nonlocal contributions due, perhaps, to the presence of Weyl tensor or to nonlocal interactions (e.g., Newtonian potential energy) between the $T_{\mu\nu}$'s at different points.

Of course this nonlocal character of gravitational energy was recognized long ago by Einstein.[4] The energy-momentum pseudo-tensor was to be regarded merely as a means to calculating the total energy-momentum. The nonlocality of gravitational energy was then exhibited in the local dependence of the pseudo-tensor on the choice of coordinate system. The numerous attempts which have since been made, aimed at removing this coordinate dependence (e.g., by means of coordinate conditions or preferred vector fields or hypersurfaces) and thus, perhaps, obtaining a meaningful local energy, have never been completely successful.[21] It also may be mentioned that, although the four-index tensor of Robinson and Bel[1] is, on the other hand, local, it does not measure energy but a higher-order moment. As has been pointed out by Trautman, [21] the nonlocality of gravitational energy need not be attributed either to the nonlinearity of the field equations or to the fact that gravity operates in a Riemannian space (although it is connected with both), since it is present already with the linear theory of a spin 2 zero-mass field. In the linear theory, the local energy density depends on potentials and cannot be given a gauge independent meaning in general. The total energy-momentum, however, is gauge invariant. Similarly, in the full Einstein theory, provided suitable asymptotic conditions at infinity are imposed on the space-time (and provided $\lambda = 0$), it is possible to define a conserved total energy-momentum independently of the coordinate gauge, and to give a

precise measure of the contributions due to radiation by means of integrations carried out entirely at infinity.[2,7,9,10,11,19] It will turn out that this definition of energy flux at large distances due to gravitation, electromagnetic, or neutrino radiation is, in fact, in essential agreement with the suggestion proposed here.

2. The Optical Scalars

Let n be a null ray (geodesic) in space-time, belonging to a congruence of null rays defined in the neighborhood of n. The convergence and rotation of the null rays near a point P of n can be combined and represented as a single complex number ρ. Similarly, the magnitude and orientation of the shear of the null rays near P are conveniently represented as a complex number σ. These are the so-called optical scalars of Robinson and Sachs.[5,17,18] (The notation of Newman and Penrose[8] is being used here.) These are not strictly scalars of the congruence, however, since the definition of ρ depends on a choice of scale on n, while σ depends on the scale and on a choice of a (null) 2-plane element tangent to n and normal to n. We can specify these by associating a real future-pointing null vector l^μ tangent to n at P, whose extent determines the scale, and a complex null vector m^μ orthogonal to l^μ, whose real part spans, with l^μ, the assigned 2-plane element at P. Let m^μ be normalized by $m_\mu \bar{m}^\mu = -1$. We may choose l^μ, m^μ at other points of n, to be given by parallel transport along n. On the other null rays of the congruence we may choose l^μ, m^μ related in the same way as on n. Then,

$$\rho = l_{\mu;\,\nu} m^\mu \bar{m}^\nu, \quad \sigma = l_{\mu;\,\nu} m^\mu m^\nu. \tag{5}$$

Alternatively, we may use the spinor approach, which is equivalent but a little more economical. A two-component spinor o^A determines both a null vector and an oriented 2-plane element related to the null vector as above. We have

$$o^A o^B \epsilon^{C'D'} = 2 l^{[\mu} m^{\nu]} \sigma_\mu{}^{AC'} \sigma_\nu{}^{BD'}, \tag{6}$$

giving the relationship between o^A and l^μ, m^μ. Then eq. (5) becomes

$$o^{A;\,BC'} o_A \bar{o}_{C'} = \rho o^B, \; o^{A;\,BC'} o_A o_B = \sigma \bar{o}^{C'}. \tag{7}$$

By defining the affine parameter r on n consistently with

$$r;_\mu l^\mu = 1, \tag{8}$$

we can relate it to a propagation derivative, D, for quantities defined on n:

$$D \equiv \frac{d}{dr} \equiv ;_\mu l^\mu. \tag{9}$$

We then have[18]

$$D\rho = \rho^2 + \sigma\bar{\sigma} + \Phi \tag{10}$$

$$D\sigma = (\rho + \bar{\rho})\sigma + \Psi, \tag{11}$$

where

$$\Phi = -\tfrac{1}{2}R_{\mu\nu}l^\mu l^\nu = \Phi_{ABC'D'}o^A o^B \bar{o}^{C'} \bar{o}^{D'} \geq 0 \tag{12}$$

and

$$\Psi = C_{\alpha\beta\mu\nu}l^\alpha m^\beta l^\mu m^\nu = \Psi_{ABCD}o^A o^B o^C o^D. \tag{13}$$

The quantities Ψ_{ABCD}, $\Phi_{ABC'D'}$ are the spinor analogues[8] of $C_{\alpha\beta\mu\nu}$ and the trace-free part of $R_{\mu\nu}$, respectively. The inequality in eq. (12) is a consequence of the definiteness of local energy in eq. (3). Note that under the allowed transformation,

$$o^A \to \lambda o^A, \tag{14}$$

or, equivalently,

$$l^\mu \to \lambda\bar{\lambda}l^\mu, \; m^\mu \to \lambda\bar{\lambda}^{-1}m^\mu, \tag{15}$$

where λ is complex and constant along n, we have

$$D \to \lambda\bar{\lambda}D, \rho \to \lambda\bar{\lambda}\rho, \sigma \to \lambda^3\bar{\lambda}^{-1}\sigma, \Phi \to \lambda^2\bar{\lambda}^2\Phi, \text{ and } \Psi \to \lambda^4\Psi. \tag{16}$$

The modulus and argument of λ determine, respectively, a change of scale on n and a rotation of the 2-plane element about n.

The geometrical meanings of ρ and σ are as follows:[18] Consider a small circle with center P lying in a spacelike 2-plane orthogonal to n. If we follow the lines of the congruence which intersect the circle, into the future null direction, the circle may become contracted (or expanded), rotated, or distorted into an ellipse. These are measured, respectively, by $\mathrm{Re}\rho$, $\mathrm{Im}\rho$ and σ (see figure 1), where $|\sigma|$ measures the magnitude of the distortion and $\frac{1}{2}\arg\sigma$ measures the angle the minor axis of the ellipse makes with the assigned null 2-plane.

The condition that

$$\rho = \bar{\rho} \tag{17}$$

at each point, since it denotes absence of rotation, is the condition for

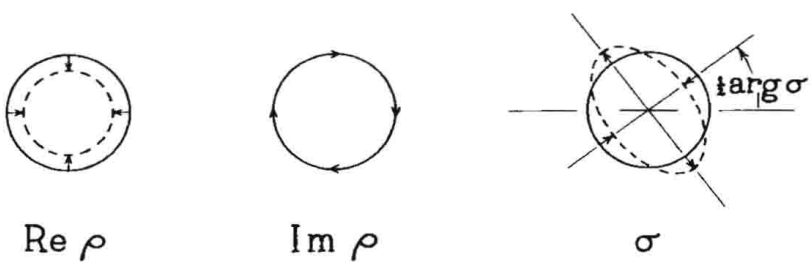


Figure 1. The geometrical interpretation of the optical scalars in terms of the effect of propagation of a small circle perpendicular to the beam.

hypersurface orthogonality of the congruence. In fact, congruences satisfying eq. (17) will be of the most interest to us here, since this also means that the null rays of the congruence generate a system of null hypersurfaces; thus, these hypersurfaces can act as wave fronts for zero rest-mass radiation. In particular, the generators of a (one-dimensional) system of null cones would give rise to eq. (17). If eq. (17) holds, then ρ and σ pertain to geometry of the individual hypersurfaces and not to the relationship between the different hypersurfaces. We may then think of the null rays near n as giving a beam of radiation. If we consider the rays belonging to a single null hypersurface only, then we have a pulse of radiation of zero duration. (The possible presence of dispersion is being ignored here. Only geometrical optics will be used.)

3. Lenses

Note that in eqs. (10) and (11) it is essentially Φ which governs the change in ρ, and it is Ψ which governs the change in σ. This is more precisely the case in the idealized situation, where Φ and Ψ are very localized and are effectively allowed to become Dirac δ-functions:

$$\Phi = \Phi_\delta \delta(r - r_0), \Psi = \Psi_\delta \delta(r - r_0). \tag{18}$$

(Idealized plane-fronted gravitational-electromagnetic waves with δ-function amplitudes can be readily constructed for which eq. (18) holds.[3,13,16] Then the values of ρ and σ will be discontinuous across $r = r_0$, with the jumps given by

$$[\rho] = \Phi_\delta \geqslant 0, [\sigma] = \Psi_\delta. \tag{19}$$

In both regions $r < r_0$ and $r > r_0$, the behavior of ρ and σ will be as in flat space. Thus, the effect of the δ-function in the curvature at $r = r_0$ on the beam is precisely the same as that of a (thin) lens. (This will be discussed in more detail in section 5.) For example, if $\Psi_\delta = 0$, we get an anastigmatic lens. The astigmatism in the beam is clearly measured by σ. Thus, if σ is unchanged as the beam passes through a lens, there will be a consequent absence of astigmatism in the lens. Similarly, if $\Phi_\delta = 0$, there will be a purely astigmatic lens; that is, the positive focusing in one plane is exactly as great as the negative focusing in the perpendicular plane. Referring to eqs. (12) and (13), we see that the effect of Ricci tensor (actually the trace-free part—the scalar curvature makes no contribution) on a beam is to introduce anastigmatic positive focusing (i.e., and increase in the convergence, since $\Phi_\delta \geqslant 0$—a magnifying lens). The effect of Weyl tensor is to produce purely astigmatic focusing.[14,18]

It is, perhaps, instructive to see how these different effects are present in the bending of light by the sun. In order to get the complete picture we must envisage the sun as transparent and without refractive index (see figure 2.) (Presumably this is the picture presented if we were to view by means of neutrinos.) Now, the line of sight passing through the center of the sun must, by symmetry, encounter no astigmatism. It encounters Ricci tensor, however, so there must be a resultant magnification of the background star field. We expect to have magnification for the other lines of sight intercepting the sun, presumably decreasing as we approach the limb, but also a certain amount of distortion due to

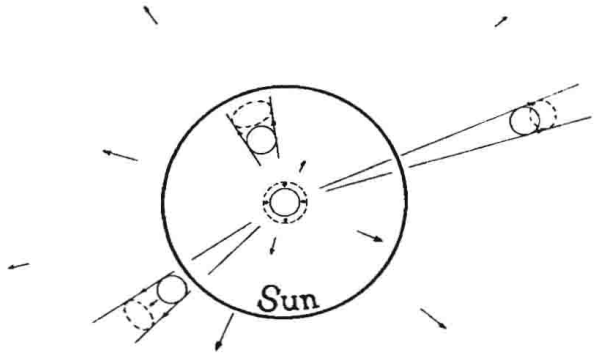


Figure 2. The effect of the sun's curvature on the appearance of the background star field. (The sun is regarded as transparent and without refractive index.) The magnification caused by Ricci tensor predominates in the central regions while, outside the sun's disc, distortion caused by Weyl tensor is evident.

Weyl tensor. In fact, the whole star field behind the sun will be magnified. As we examine regions farther and farther from the limb, however, we will get a decrease in the displacement caused by the magnification of the star field. A circular pattern in the star field will be displaced outward by a smaller amount at its part farthest from the center of the sun than it will at its part nearest to the center. On the other hand, since the whole circle is to move outward and remain tangent to two fixed lines through the sun's center, it must become stretched in the perpendicular direction. The small circle thus becomes elliptical—the effect of astigmatic focusing due to the Weyl curvature of the sun's gravitational field!

It appears, then, that the focusing effect due to local stress-energy and the effect due to local gravitational (Weyl) field are of totally different kinds (and are, in a sense, "orthogonal" to each other). There is, however, more to the matter than this, since these considerations have so far been concerned only with the ("first order") effect of local focusing. To get a complete picture, we must take into account the nonlinearity of eqs. (10) and (11). Thus, the resultant effect of a series of lenses intercepted by a beam is not simply the sum of the individual effects due to each lens separately.

Consider, for simplicity, the case of two lenses only. In elementary optical theory, the power of a lens is the reciprocal of its focal length (positive for a converging lens). An astigmatic lens will have different powers for each plane perpendicular to the lens, but in an anastigmatic lens all these powers are equal and simply measure the strength of the lens. The power is additive for lenses placed directly against one another, but if two (anastigmatic) lenses of respective powers p_1 and p_2 are placed a distance w apart, the resulting power of the combined system is given by*

$$p = p_1 + p_2 - p_1 p_2 w. \tag{20}$$

---

* See, for example, reference 20, p. 108.

Actually the situation is complicated slightly by the fact that the resulting system behaves not as a thin lens but as a thick lens. A thick anastigmatic lens is characterized by two numbers; namely, by its focal length (or by its power) and by the distance k between its principal planes.[20] The case $k = 0$ corresponds to a thin lens and in this case the principal planes coincide with the lens. The effect of a general thick lens differs from that of the corresponding thin lens only in that the rays, at the point of deflection (at a principal plane), also are displaced a distance k (i.e., from the one principal plane to the other). In fact, it is a classical theorem of elementary optics that any system of perfect anastigmatic lenses is equivalent in its resultant effect to a single anastigmatic thick lens (with positive or negative k)—provided we allow certain limiting cases of a thick lens to count as a thick lens. For the above situation we have, in fact,

$$k = -\left(\frac{p_1 p_2}{p}\right) w^2. \tag{21}$$

The case of astigmatic lenses is more complicated, but in simple cases it can be treated in essentially the same way. Consider the case of two equal and opposite purely astigmatic lenses. (Thus, each one is the same as the other but rotated through a right angle in the plane of the lens.) When placed directly against each other, they cancel out to produce zero effect. Now move them apart so that the distance between them is w. We can apply eq. (20) in each of two planes perpendicular to the lenses, namely, the planes of maximum (positive or negative) power $\pm p_0$. In one of these planes we have, for example, $p_1 = p_0, p_2 = -p_0$; in the other we have $p_1 = -p_0, p_2 = p_0$. In either case eq. (20) gives

$$p = p_0^2 w \tag{22}$$

for the resulting power. Thus, we apparently get an anastigmatic lens of power $p_0^2 w$. From eq. (21), the distance between the principal planes is given by $k = w$ in each case, and it might appear, at first sight, that we have an ordinary thick anastigmatic lens. Actually this is not quite the case, because the position of the principal planes is not the same for all rays; they vary over a total distance $2p_0^{-1}$. If w is small, however, compared with the focal distance p_0^{-1}, then $2p_0^{-1}$ is small compared with p^{-1}, and $k(=w)$ is very small compared with p^{-1}. Thus, when viewed on a sufficiently large scale (that is, sufficiently larger than p_0^{-1}), the focusing effect of the system will appear to approximate that of an anastigmatic positive lens of power $p_0^2 w$. This behavior seems to be quite general, as we shall see presently.

4. Energy Flux as Focusing Power?

Let us see if this behavior is consistent with an interpretation of energy flux across a null ray n as the focusing power exerted by the curvature along n. We have seen that focusing power is essentially additive over distances which are small compared with the focal lengths involved and that, locally, the net focusing power of a purely astigmatic lens is zero. These observations correspond to the fact that, over distances which are small compared with the radii of curvature involved in $R_{\mu\nu\alpha\beta}$, we can

treat energy as additive and the contribution of the Weyl tensor to the energy as essentially zero. We see, indeed, that focusing power is measured locally by Φ, which is given by the flux $T_{\mu\nu}l^\mu l^\nu$ of local energy across n—see eqs. (3) and (12); the Weyl tensor contributes locally only to the astigmatism and not to the focusing power.

When larger distances are involved, however, or a higher degree of accuracy is required, the nonlinear (or nonlocal) effects must be taken into account. In fact, eq. (20) does this in a simple case. Suppose the null ray n intercepts two localized patches of Ricci curvature separated by an affine parameter distance w. The resultant energy flux will not be simply the sum of the two contributions, but will be corrected by what amounts, effectively, to the Newtonian potential energy of interaction between the masses of the two patches. The negative sign of the correction term in eq. (20) agrees with the fact that the Newtonian potential energy is negative. The product of the powers of the two lenses is involved, in agreement with the Newtonian "product of the masses." Moreover, the dependence on distance appearing in eq. (20) is what we should expect from an interpretation as a Newtonian potential. We are, in effect, dealing with a one-dimensional situation, because only a single ray is involved. For example, if we imagine the situation which arises when two electromagnetic (approximately) plane waves of short duration are intercepted by n, the Newtonian potential will be proportional (for small distances) to the distance between the waves because we have (approximate) plane symmetry. Compare this result with eq. (20). If, on the other hand, we have two localized masses producing the Ricci curvature, the number of null rays intercepting both masses goes roughly inversely as the square of the mass separation. If we use eq. (20) for the null rays which do intercept both masses, we seem to have the correct general order of behavior for a Newtonian "reciprocal of distance" potential.

Let us consider now the free gravitational case. Again it will be convenient to consider plane or approximately plane waves intercepted by n. Suppose we have a (nonoscillatory) gravitational pulse followed by an equal and opposite gravitational pulse after an interval w. When $w = 0$, the two pulses cancel out and the energy intercepted by n is zero. When w is positive but still small compared with the reciprocals of the strengths of the two pulses, consider the focusing effect, and refer to the discussion at the end of the preceding section; the pair of pulses will appear on a large scale to be essentially the same as a single pulse of matter (e.g., electromagnetism) whose strength is given by the product of w with the square of the strength of the gravitational pulses. Refer again to eq. (22). Thus, it seems reasonable to identify the energy carried by the pair of gravitational pulses as (at least approximately) the same as that which would be carried by such an equivalent matter pulse.

There is a feature of this situation which may seem surprising at first. While the curvature is situated at the pulses, we can apparently interpret the energy as "residing" in the flat space between the two pulses. For, by simply altering the interval between the two pulses and not essentially altering the pulses themselves, we can change the amount of energy flow to any value we like—and this energy will be proportional to the interval between the pulses. Other interpretations of the "location" of the gravitational energy are also possible. For example, we could

say that the first pulse contains no energy, but its presence affects the energy content of the second pulse in such a way that it produces the required amount. In fact, such different possible interpretations for the "location" of the energy already can be made in the linear theory of a spin 2 zero-mass field. The "energy density" may be expressed either as a quadratic expression in an integral of the linearized curvature field (so that "local energy flux" may be present in field-free regions), or bilinearly in the curvature field and a second integral (i.e., the linearized gauge dependent metric) of the curvature field (so the energy flux "at" one point could depend on field values at distant points). These two possible types of interpretation also will emerge in section 5 when we consider some more exact expressions of focusing power in the full gravitational theory.

Before doing this, however, one further rough argument will be given which serves to emphasize the general correctness of this behavior for gravitational energy. Consider the energy carried by a gravitational (plane-polarized) wave packet which we shall imagine to be roughly approximated by a series of pulses of Weyl curvature of the type considered above and intercepted by n. We have, therefore, a series of purely astigmatic lenses on n, each of which is essentially equal and opposite to the preceding one and separated by an affine distance w on n. Let Ψ_δ represent the strength of each pulse, with Ψ_δ as in eq.(18). The amplitude Ψ of the smoothed-out wave packet then will be of the order

$$\Psi \sim \Psi_\delta w^{-1}. \tag{23}$$

If we combine the lenses in pairs and use eq. (22), we get

$$|\Psi_\delta|^2 w \tag{24}$$

for the strength of the anastigmatic lens representing each pair. (This identification of $\pm|\Psi_\delta|$ with the principal powers of a purely astigmatic lens is justified in the next section.) Now suppose that the wave train is essentially N periods long. Thus, a total of 2N lenses will be used to represent it. Suppose also that $(|\Psi_\delta|^2 w)^{-1}$ is large compared with wN (i.e., $wN^{1/2} \ll |\Psi_\delta|^{-1}$) so that the wave is sufficiently weak or of high frequency and the wave train is not too long. Then we can ignore the correction term in eq. (20) and simply add the powers of the anastigmatic lenses linearly. The total power p for the lens effect of the entire wave packet then is

$$p \sim N|\Psi_\delta|^2 w \sim N|\Psi|^2 w^3. \tag{25}$$

Thus, identifying p with the energy flux across n, we see that the energy of a wave packet of fixed extent (wN) and fixed width is proportional to the square of the amplitude and to the square of the wavelength. This may be contrasted with the electromagnetic case, where the corresponding argument applied to a sufficiently weak, and not too long, wave train would give

$$p \sim N|\phi|^2 w, \tag{26}$$

where ϕ measures the (complex) amplitude of the electromagnetic field. Thus, the energy of a wave train of extent wN is independent of the wavelength. The amplitude ϕ is given by

$$\phi = F_{\mu\nu} l^{\mu} m^{\nu} = \phi_{AB} o^{A} o^{B}, \tag{27}$$

in suitable units, where the (skew-symmetric) tensor $F_{\mu\nu}$ or the (symmetric) spinor ϕ_{AB} represents the electromagnetic field (related by $F_{\mu\nu}\sigma^{\mu}{}_{AP'}\sigma^{\nu}{}_{BQ'} = \phi_{AB}\epsilon_{P'Q'} + \epsilon_{AB}\bar{\phi}_{P'Q'}$). We have, in this case, $\Phi = (2)\phi\bar{\phi}$. In the case of neutrinos, we have Φ of the form $\Phi = i\nu D\bar{\nu} - i\bar{\nu}D\nu$, where ν is the complex amplitude of a (positive frequency) Weyl neutrino field ν_A:

$$\nu = \nu_A o^A. \tag{28}$$

Hence, the analog of eqs. (25) and (26) gives

$$p \sim N|\nu|^2, \tag{29}$$

so that the energy of a wave train of extent Nw is proportional to the square of the amplitude and to the reciprocal of the wavelength. Equations (25), (26), and (29) are consistent with a general pattern,

$$p \sim N|\chi|^2 w^{2s-1}, \tag{30}$$

for a zero-mass field of spin s, given by a symmetric 2s-index spinor $\chi_{AB\ldots L}$, the amplitude χ being

$$\chi = \chi_{AB\ldots L} o^A o^B \ldots o^L. \tag{31}$$

There are serious difficulties in incorporating zero-mass fields of spin greater than unity into general relativity, with the notable exception of gravitation itself, because of the nonlocality of their energy. In any case, eq. (30) has relevance to spins 0, $\frac{1}{2}$, 1, 2.

The dependence on w given by eq. (30), and, therefore, also by eq. (25), is in fact a necessary consequence of the four-vector nature of the energy-momentum of a wave packet (assuming the curvature to be small so that special relativity effectively applies). If we represent the energy-momentum of the wave packet as Ek_μ, with k_μ null and normalized by $l^\mu k_\mu = 1$, then, under the transformation, eq. (15), we get $k_\mu \to |\lambda|^{-2} k_\mu$, from which

$$E \to |\lambda|^2 E. \tag{32}$$

Also, since wl^μ is an absolute quantity invariant under eq. (15), we have $w \to |\lambda|^{-2} w$. Substituting eq. (14) into eq. (31), we get $\chi \to \lambda^{2s}\chi$; from this result, eq. (30) gives

$$p \to |\lambda|^2 p, \tag{33}$$

in agreement with eq. (32). Any other power of w appearing in eq. (25) or (30) would have been inconsistent with eq. (32).

5. Focusing Power of a General System

The preceding arguments all have been concerned with focusing when the effective lenses are in some sense weak. The connection between energy flux and focusing power is, perhaps, not so clear when the lenses are strong. But, in any case, the whole concept of energy flow in general relativity becomes obscure for regions where space-time deviates too much from flatness. On the other hand, the concept of focusing power can be made quite precise. There is some hope, therefore, that this concept may be of use in clarifying the energy concept.

In order to define the focusing power of a general system along a ray n, we proceed as follows: Let us suppose, in the first instance, that the system is finite in extent along n; that is to say, both Φ and Ψ vanish if r lies outside a certain finite range. Then we can consider an incident parallel beam (or pulse—see section 2) and examine the effect of the system on the emergent beam. The general idea is to find quantities which measure correctly the focusing power when the system consists of a single lens and to use these same quantities to define the focusing power of a general system. In general, both the shear and convergence will be nonzero for the emergent beam, while they vanish for the incident beam. The presence of shear in the emergent beam is a complication and some care is needed in finding the correct definitions. Apparently, it beomes necessary to consider, in addition to ρ and σ, quantities which measure the size, ellipticity, and degree of rotation of the emergent beam, as compared with the incident beam which is assumed to have a circular cross section of unit (infinitesimal) radius.

Let ρ and σ be as set forth in section 2, and introduce two complex quantities, ξ, and η, defined along n by

$$D\xi = -\rho\xi - \sigma\bar{\eta} \tag{34}$$

$$D\eta = -\rho\eta - \sigma\bar{\xi}. \tag{35}$$

It is convenient to introduce the matrix notation

$$\mathbf{X} = \begin{pmatrix} \xi & \eta \\ \bar{\eta} & \bar{\xi} \end{pmatrix}, \mathbf{P} = \begin{pmatrix} \rho & \sigma \\ \bar{\sigma} & \bar{\rho} \end{pmatrix}, \mathbf{Q} = \begin{pmatrix} \Phi & \Psi \\ \bar{\Psi} & \Phi \end{pmatrix}. \tag{36}$$

Then eqs. (10), (11), (34), and (35) become, respectively,[18]

$$D\mathbf{X} = -\mathbf{P}\mathbf{X} \tag{37}$$

$$D\mathbf{P} = \mathbf{P}^2 + \mathbf{Q}. \tag{38}$$

From the above, we see that

$$D^2\mathbf{X} = -\mathbf{Q}\mathbf{X}. \tag{39}$$

The respective quantities $|\xi|$ and arg ξ can be used, roughly speaking, to describe linear size and degree of rotation of a beam at a point of n when η is small, while $|\eta\xi^{-1}|$ and $\frac{1}{2}$ arg $(\eta\xi)$, respectively, describe degree of ellipticity and orientation of ellipticity of a cross section of the

General-relativistic energy flux and elementary optics. 1966 *Perspectives in Geometry and Relativity (essays in Honor of Václav Hlavatý). Ed. Banesh Hoffmann pp. 259–274*

beam. If we start with a beam at $r = 0$ for which $X = I$ (the unit matrix)—the case of circular cross section, unrotated and of unit infinitesimal radius—and apply eq. (37) when P is taken to be constant, so that $X = e^{-rP} \doteq I - rP$ for small r, these interpretations of ξ and η then follow roughly from those of ρ and σ depicted in figure 1. More precisely, the ellipse of a cross section of the beam is that described in the Argand plane by $e^{i\theta}\xi + e^{-i\theta}\eta$ as θ varies from 0 to 2π. The major and minor axes, respectively, are given by $|\xi| + |\eta|$ and $|\xi| - |\eta|$. (The ellipse is "flat" if $|\eta| = |\xi|$ and "inside-out" if $|\eta| > |\xi|$. If $\xi = 0$, the beam again has a circular cross section, but with a negative orientation. In fact, there is an effective reversal of the roles of ξ and η when $|\eta| > |\xi|$.) The area of cross section of the beam is

$$a = \pi(\xi\bar{\xi} - \eta\bar{\eta}) = \pi|X|. \tag{40}$$

Let us now consider the case of a thin lens situated at $r = 0$, given by eq. (18) (with $r_0 = 0$):

$$Q = Q_\delta \delta(r). \tag{41}$$

For the incident beam we take $X = I$, $P = 0$ so (see eqs. (19) and (37)),

$$X = I, -DX = P = Q_\delta \text{ at } r = 0+. \tag{42}$$

Since, by eq. (39) we have $D^2X = 0$ for $r > 0$, it follows (see eq. (42)) that

$$X = I - rQ_\delta, (r > 0). \tag{43}$$

Note that, by eq. (37), $-Q_\delta = -P(I - rQ_\delta)$, from which we see that

$$P = Q_\delta(I - rQ_\delta)^{-1}, (r > 0). \tag{44}$$

Also, by eqs. (40) and (43),

$$a = \pi|I - rQ_\delta|, (r > 0). \tag{45}$$

The focal points are given by $a = 0$, i.e., when r^{-1} is an eigenvalue of Q_δ (see also eq. (44)). These eigenvalues are $\Phi_\delta \pm |\Psi_\delta|$, a fact which justifies the interpretation of

$$\Phi_\delta + |\Psi_\delta|, \text{and } \Phi_\delta - |\Psi_\delta| \tag{46}$$

as the maximum and minimum powers of the lens, as was implicit in the discussion of section 4.

We wish to obtain the powers, eq. (46), by examining the emergent beam, but without using knowledge of the position of the lens. Then the same technique could be applicable to more general systems. Thus, we must express Q_δ in terms of X and P but independently of r. This is achieved by using eqs. (43) and (44):

$$Q_\delta = PX, (r > 0). \tag{47}$$

To express the total energy flux across the beam, it is presumably Φ_δ that is required; namely

$$\rho\xi + \sigma\bar{\eta}, \tag{48}$$

while $\rho\eta + \sigma\bar{\xi}$ measures the resultant effective astigmatism (see eq. (35)). We can thus envisage also using eq. (48) in more general situations than just that of a single lens, and define the total energy flux across n, up to a certain point of n, as eq. (48) evaluated at that point.

A number of features of eq. (48) may be mentioned. In the first instance it is not necessarily real for a general system. The phase of eq. (48) is, in fact a measure of the amount of resultant rotation of the beam. (A rotation of the beam results if, for example, n intercepts a circularly polarized gravitational wave.) Thus, a possibility would be to replace eq (48) as a definition of energy flux by its modulus, or perhaps by its real part. In view of the fact that a complex expression—whose integral turns out to be real—occurs naturally in an asymptotic formula for total energy-momentum, however,[10,11] it would seem not unreasonable to use eq. (48) as it stands.

One interesting and satisfying property of eq. (48) is that the measure of focusing power it gives is the same (actually complex conjugate) as for the reverse situation, when it is the emergent beam which is parallel with unit circular cross section. To see this, consider the expression

$$V_{12} = X_1{}^*(P_1{}^* - P_2)X_2 \tag{49}$$

where the indices 1, 2 refer to two different beams along the line n. The asterisk denotes hermitian conjugate. The matrix, V_{12}, is then constant along n:

$$DV_{12} = 0, \tag{50}$$

since $D\{(DX_1)^*X_2 - X_1{}^*(DX_2)\} = (D^2X_1)^*X_2 - X_1{}^*(D^2X_2) = -(QX_1)^*X_2 - X_1{}^*(QX_2) = 0$, because $Q^* = Q$. Now let the quantities with suffix 1 refer to the beam of the original situation so that $X_1 = I$, $P_1 = 0$ initially, and, therefore, $V_{12} = -P_2X_2$ initially. Similarly, let the quantities with suffix 2 refer to the situation in which the emergent beam ultimately satisfies $X_2 = I$, $P_2 = 0$. Thus $V_{12} = X_1{}^*P_1{}^*$. Hence the initial value of $\rho_2\xi_2 + \sigma_2\bar{\eta}_2$ is equal to minus the complex conjugate of the final value of $\rho_1\xi_1 + \sigma_1\bar{\eta}_1$. (A similar result holds for the effective astigmatism measure $\rho\eta + \sigma\bar{\xi}$, but without the complex conjugation—the change of sign is to be expected, since the focal lengths are measured in the opposite direction.)

When the ellipticity vanishes, we have $a = \pi|\xi|^2$, so the modulus of eq. (48) can be written in another form:

$$\rho a^{1/2}\pi^{-1/2}. \tag{51}$$

This has some interest because the derivative (see eqs. (10), (38), and (40))

$$D(\rho a^{1/2}) = a^{1/2}(\Phi + \sigma\bar{\sigma}) \tag{52}$$

is always nonnegative when $a > 0$. In circumstances when it is fair to ignore the ellipticity (e.g., for suitable rapidly oscillating fields), this seems to reflect a physically desirable positive definiteness of "local" energy flow. The status of the derivative of eq. (48) is not so clear, however:

$$D(\rho\xi + \sigma\bar{\eta}) = \Phi\xi + \Psi\bar{\eta} \tag{53}$$

(see eqs. (37) and (39)). The Weyl tensor contribution, $\Psi\bar{\eta}$ has the rough form of "curvature field times second integral of curvature field," whereas in eq. (52) the contribution, $\sigma\bar{\sigma}$, is essentially "absolute square of integral of curvature field." These two types of interpretation for the "local" gravitational energy flux are, in fact, essentially the two possibilities envisaged in section 4. Thus, eq. (52) may be nonzero in field-free regions, whereas the effect of Ψ in eq. (53) is governed by the ellipticity of the beam, i.e., by the nonlocal effects of Ψ at distant points.

Apparently because of its positive definite nature, it is essentially the asymptotic form of eq. (52) rather than that of eq. (53) which has emerged[2,7,9,10,11,19] as the definition of energy flow at infinity due to electromagnetic, neutrino, or gravitational radiation from an isolated system. In an asymptotically flat space-time the null ray n may be moved out towards infinity until in the limit it becomes a "generator of the null hypersurface at infinity."[10,11] The quantity Φ becomes, in the limit, the energy flux at infinity due to electromagnetic radiation (the absolute square of the radiation field) plus the energy flux due to neutrino radiation (a bilinear expression in the radiation field and its time derivative). The quantity σ becomes the Bondi-Sachs "news function," (a time integral of the gravitational radiation field) whose absolute square is taken to represent the energy flux at infinity due to gravitational radiation.

Although the value of the "news function" is well defined when the space-time is known as a whole, it is, as might be expected from the foregoing discussion, still essentially a nonlocal quantity. That is, complete knowledge of the geometry of space-time in the neighborhood of part of infinity will be insufficient for determination of the "news function" there. Its precise determination depends on a detailed knowledge of the asymptotic geometry along a complete "cone" of null directions, or else on the complete history in one null direction.

In view of the definiteness of eq. (52) and of its relationship to the above asymptotic energy flux definitions, it is tempting to ask whether $\rho a^{1/2}$ might not better measure total energy flux than the complex quantity $\rho\xi + \sigma\bar{\eta}$. A serious objection to its use, however, results from the fact that $\rho a^{1/2}$ becomes infinite at any astigmatic focal point of the beam. (It becomes purely imaginary just beyond such a point.) The existence of such focal points somewhere along a beam is, in fact, the normal situation, and it seems unreasonable to have to assign an infinite value for the gravitational energy flow there. Indeed, we have (with $\rho = \bar{\rho}$)

$$D(a^{1/2}) = -\rho a^{1/2}, \tag{54}$$

from which, by eq. (52), $D^2(a^{1/2}) < 0$ (assuming Φ, Ψ not both identically

zero), so that a focal point must eventually be encountered by n. This is the effect described by Raychaudhuri[15] as applied to null rays.* The quantity $\rho\xi + \sigma\bar{\eta}$, on the other hand, never becomes infinite in a nonsingular space-time.

As for the geometrical interpretation of $\rho\xi + \sigma\bar{\eta}$, we see from eq. (34) that it is the propagation derivative of $-\xi$. It is thus minus the rate of change of the "size" and "orientation" of the beam, where "size" is given by the arithmetic mean of the major and minor axes of cross section. On the other hand, by eq. (54), $\rho a^{1/2}$ is minus the rate of change of $a^{1/2}$, which is $\pi^{1/2}$ times the geometric mean of the axes of cross section. It therefore becomes evident why, in flat space-time, ξ varies linearly with r, whereas $a^{1/2}$ generally does not.

It may be asked whether ξ itself has any simple physical interpretation. There is some indication, in fact, that the value of ξ at a point R of n (for a beam where initially $X = I$, $P = 0$) may be regarded as a measure of the gravitational potential at R due to the total energy intercepted by n up to R (i.e., the retarded contribution along n at R). Consider a portion of the null cone of R consisting of a set of generators neighboring n (i.e., a pulse) for which $P_R X_R \to I(\rho_R \to \infty, \sigma_R \to 0)$ at R. (A suffix R distinguishes quantities associated with this pulse.) Then eqs. (49) and (50) tell us that the value of X^* at R is equal to the limiting value of $-P_R X_R$, in the past along n. Thus, the value of $-(\rho_R\xi_R + \sigma_R\bar{\eta}_R)$ in the distant past along n is the same as the value of $\bar{\xi}$ at R. The dependence of $D(\rho_R\xi_R + \sigma_R\bar{\eta}_R) = \Phi\xi_R + \Psi\bar{\eta}_R$ on distance and on the mass intercepted by the pulse (taking $\Psi = 0$) is $\sim (\Phi r^2)r^{-1}$ (since $\xi_R \sim r$), which is consistent with the above interpretation of it as a potential and, in fact, it is also consistent with an interpretation of $\rho_R\xi_R + \sigma_R\bar{\eta}_R$ at a point S (to the past of R on n) as a measure of the contribution to the potential at R from the segment SR. It is possible to relate this finding roughly to the behavior of clocks, but the matter will not be entered into here.

Since $-D\xi_R = \rho_R\xi_R + \sigma_R\bar{\eta}_R$, we might also ask whether ξ_R has any simple physical interpretation. No suggestion will be made here except to mention that it is an analog of the luminosity distance $a_R^{1/2}$ from R. It possesses one interesting property; namely, the symmetry that ξ_R at S equals $-\bar{\xi}_S$ at R, as readily follows from eqs. (49) and (50). Also, η_R may be of interest physically. It possesses a similar symmetry. Note the fact that this and eq. (40) imply that the luminosity distance from R to S is the same as that from S to R. (A suggestion that this might be the case made to me earlier by R. K. Sachs.†) The result also follows more directly by using the determinant of eq. (49). All these results have (nontrivial) analogs in optical theory.

REFERENCES

1. Bel, L.: Compt. Rend., **248**: 1297 (1959).

2. Bondi, H., van der Burg, M. G. J., and Metzner, A. W. K.: Proc. Roy. Soc., **A269**: 21 (1962).

---

* See also reference 12.

† See also reference 6.

3. Brinkman, H. W.: Proc. Nat. Acad. Sci., **9**: 1 (1923).
4. Einstein, A.: S. B. preuss. Akad. Wiss., 488 (1918).
5. Jordan, P., Ehlers, J., and Sachs, R. K.: Akad. Wiss. Lit. Mainz. A Math.-Nat. Kl., **1** (1961).
6. Kristian, J., and Sachs, R. K: Astrophys. J., **143** : 379 (1966).
7. Morgan, T., and Winicour, J : Proc. Amer. Phys. Soc. (Washington 424 (1964).
8. Newman, E. T., and Penrose, R.: J. Math. Phys., **3**: 566 (1962).
9. Newman, E. T., and Unti, T. W. J.: J. Math. Phys., **3**: 896 (1962).
10. Penrose, R.: Phys. Rev. Let., **10**: 66 (1963).
11. Penrose, R.: in Relativity Groups and Topology, DeWitt, ed., 196
12. Penrose, R.: Phys. Rev. Let., **14** : 57 (1965).
13. Peres, A.: Phys. Rev. Let., **3**: 571 (1959).
14. Pirani, F. A. E., and Schild, A.: Bull. Acad. Polon. Sci. Math. Astr Phys., **9**: 543 (1961).
15. Raychaudhuri, A.: Phys. Rev., **98**: 1123 (1955).
16. Robinson, I.: Report to Eddington Group, Cambridge, 1956.
17. Robinson, I.: J. Math. Phys., **2**: 290 (1961).
18. Sachs, R. K.: Proc. Roy. Soc., **A264**: 309 (1961).
19. Sachs, R. K.: Proc. Roy. Soc., **A270**: 103 (1962).
20. Sears. F. W.: Optics, 3rd ed, Reading, Mass.: Addison-Wesley, 19
21. Trautman, A.: in Gravitation: An Introduction to Current Resear p. 169 (Witten, ed.) New York: John Wiley & Sons, 1962.

26
NOTE ON THE BONDI–METZNER–SACHS GROUP

J. Math. Phys. **7(5)** (1966), 863–870. With E. T. Newman

In this paper we showed how information at spacelike infinity i^0 can be used to reduce the BMS freedom to that of the Poincaré (or inhomogeneous Lorentz) group. Unfortunately, this does not help very much with the problem of defining angular momentum at some later retarded time. We also introduced the useful operator $\eth$, which acts on spin-weighted functions on the 2-sphere; see also Chapter 58.

Note on the Bondi–Metzner–Sachs Group*

E. T. NEWMAN
University of Pittsburgh, Pittsburgh, Pennsylvania
AND
R. PENROSE
Birkbeck College, University of London, London, England
(Received 28 October 1965)

It is shown that, in space–times which are asymptotically flat, there are reasonable physical restrictions that allow one to impose coordinate conditions (in addition to the usual Bondi–type conditions) which restrict the allowed coordinate group to a subgroup of the Bondi–Metzner–Sachs group. This subgroup is isomorphic to the improper orthochronous inhomogeneous Lorentz group.

I. INTRODUCTION

In recent years the subject of gravitational radiation has received a great deal of attention.[1–5] One of the unusual results of this study was that, even in space-times which are asymptotically Minkowskian, one apparently cannot extract the inhomogeneous Lorentz group as an asymptotic symmetry group if finite (retarded) time intervals, only, are considered. Instead, one obtains what is known as the Bondi–Metzner–Sachs group (BMS group)—an infinite parameter group. The purpose of this note is to show that if one imposes certain apparently reasonable physical restrictions at retarded time $u = -\infty$ (or, alternatively, at $u = +\infty$), it is possible to introduce further coordinate conditions such that a well-defined (non-normal) subgroup of the BMS group—isomorphic to the improper orthochronous inhomogeneous Lorentz group—is geometrically singled out by preserving these coordinate conditions.

The physical situations we allow would appear to include a general type of scattering problem in which the sources, and perhaps some waves, can come in from infinity and can again escape to infinity after interacting. This generalizes certain situations considered by Sachs. He observed[6] that the inhomogeneous Lorentz group could be singled out at, say, $u = +\infty$ if, for example, all the matter were radiated away as zero rest-mass energy, leaving Minkowski space for sufficiently large u; alternatively, the system might ultimately simply become static. The situations we consider here appear to be much more general than this, although whether or not they are quite as general as they seem to be, depends, to some extent, on the validity of some heuristic arguments we give, which have, as their basis, some exact results in the linear theory of gravitation.

As an additional purpose of this note, we introduce an invariant differential operation on the sphere (which we denote by $\eth$) and use it to define a type of "spin s spherical harmonic." This is done in slightly greater generality than is absolutely necessary for the present work, since it is also felt that these ideas should find applications elsewhere[7] and that they could be used to simplify earlier work.[2,4,6,8]

II. THE BMS GROUP

Let u, θ, ϕ, r be standard (Bondi-type) coordinates[1,2,4,6] for asymptotically flat space–time. Thus u is a *retarded time* parameter (so that u = const are null hypersurfaces opening into the future); θ, ϕ are spherical polar coordinates for the sphere at infinity on each hypersurface u = const (θ, ϕ, u = const giving the null geodesic generators of these hypersurfaces); r is suitably defined radial coordinate such as an affine or luminosity parameter on each of the generators of the hypersurfaces. The BMS group is defined by the following transformations[2,4,6] on the θ, ϕ, u coordinates:

$$\begin{aligned} \theta' &= \theta'(\theta, \phi), \\ \phi' &= \phi'(\theta, \phi), \\ u' &= K(\theta, \phi)\{u - \alpha(\theta, \phi)\}, \end{aligned} \tag{2.1}$$

where $(\theta, \phi) \to (\theta', \phi')$ is a conformal transformation

* Work supported, in part, by the Aerospace Research Laboratory through the European Office of the U. S. Air Force (and, in part, by Contract AF–AFOSR–454–63).

[1] H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, Proc. Roy. Soc. (London) **A269**, 21 (1962).

[2] R. K. Sachs, Proc. Roy. Soc. (London) **A270**, 103 (1962).

[3] E. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962); **4**, 998 (1963).

of the (θ, ϕ)-sphere into itself, K is the corresponding conformal factor, given by

$$d\theta'^2 + \sin^2\theta'\, d\phi'^2 = K^2(d\theta^2 + \sin^2\theta\, d\phi^2), \qquad (2.2)$$

and where α is an arbitrary (suitably smooth) real function on the sphere. (The r coordinate may also be transformed, if desired, but the transformation is somewhat arbitrary since it depends on the precise type of radial coordinate used and it is not relevant to the structure of the group.)

The particular BMS transformations for which $\theta' = \theta, \phi' = \phi$ are called, in general, *supertranslations*. Under a supertranslation, the system of null hypersurfaces u = const. is transformed into a different system of null hypersurfaces (u' = const) but no (Lorentz) rotation is involved. We may write α in terms of spherical harmonics:

$$\alpha = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{m,l} Y_{l,m}(\theta, \phi),$$

where the $a_{m,l}$ are constants. The *infinite set* of parameters $a_{m,l}$ [subject to $a_{-m,l} = (-1)^m \bar{a}_{m,l}$, so that α is real] then define the supertranslation. If $a_{m,l} = 0$ for $l > 2$, so that α takes the form:

$$\alpha = \epsilon_0 + \epsilon_1 \sin\theta\cos\phi + \epsilon_2 \sin\theta\sin\phi + \epsilon_3 \cos\theta, \qquad (2.3)$$

then the supertranslations reduce to a special case, called the *translations*, with just *four* parameters $\epsilon_0, \cdots, \epsilon_3$.

This terminology is, in fact, consistent with that for the (Lorentz) translations in Minkowski space. We may call a hypersurface u = const a "good" cone in Minkowski space if it is the null cone of some point, and a "bad" cone if, on the other hand, the generators of u = const do not all meet in a point. Thus, an actual translation in Minkowski space must send "good" cones into "good" cones; under a general supertranslation, the "good" cones will be warped into "bad" cones. It is precisely the condition that α be given by (2.3), which is required to preserve the "goodness" of the Minkowski null cones.

In curved asymptotically flat space–times the difficulty is to find an appropriate analog of the Minkowskian concept of "good" and "bad" cones. It is not, in fact, necessary to do this in order to single out the translations from the remaining supertranslations, since the translations are already determined by (2.3).[9] But if we wish, in addition, to isolate the "pure" Lorentz *rotations* from Lorentz rotations which have a "supertranslation component," then some concept of a distinction between "good" and "bad" cones is necessary. For we might try to define Lorentz rotations (homogeneous Lorentz transformation) as given by (2.1) with $\alpha = 0$. The hypersurface $u = 0$ is then transformed into itself. In Minkowski space, if $u = 0$ is a "good" cone, the resulting transformation indeed represents a Lorentz rotation and it sends other "good" cones into "good" cones. If, on the other hand $u = 0$ is a "bad" cone, then we do *not* get a Lorentz rotation in general. Thus, for asymptotically flat spaces, in order to know which of the BMS transformations are to be regarded as "supertranslation-free Lorentz rotations", we must have some definition of "goodness" of u = const hypersurfaces.

In Minkowski space, the "good" cones can be characterized locally by the fact that the null rays generating them possess no shear. In asymptotically Minkowskian spaces, it will only be the *asymptotic* behavior of the (complex) shear σ of these null rays that will concern us. With null vector l^μ tangent to the null rays and complex null vector m^μ orthogonal to l^μ satisfying $m^\mu \bar{m}_\mu = -1$, l^μ and m^μ being parallelly propagated along each ray, we have[3]

$$l_{\mu;\nu} m^\mu \bar{m}^\nu = \sigma(u, \theta, \phi, r) = \sigma^0(u, \theta, \phi)/r^2 + O(r^{-4}), \qquad (2.4)$$

where r is scaled so that $r, {}_\mu l^\mu \to 1$ at infinity. Thus σ^0 defines the asymptotic shear of the hypersurface given, say, by u = const. The complex quantity σ^0 is of special interest in gravitational radiation theory. It forms part of the initial data on $u = 0$ used to determine the space-time asymptotically.[2,4] Furthermore, $\partial\sigma^0/\partial u$ and $\partial^2\sigma^0/\partial u^2$ both have physical significance. We may call $\partial^2\sigma^0/\partial u^2$ the *gravitational radiation field* since it represents the r^{-1} part of the Riemann curvature field.[1,2,4] Bondi *et al*[1] and Sachs[2] call $\partial\sigma^0/\partial u$ the "news function" since it can be used as asymptotic initial data for the gravitational radiation field and $|\partial\sigma^0/\partial u|^2$ represents the flux of energy of the gravitational radiation in their analysis.

We cannot, however, attempt to define "good" cones, in general, simply by requiring $\sigma^0 = 0$. In many cases it is simply not possible to arrange $\sigma^0 = 0$ for all values of θ, ϕ, on one hypersurface, but even in cases where it is possible (e.g. in the axially symmetric, reflection symmetric cases), it is clear from the above remarks that, in the presence of gravitational radiation, if $\sigma^0 = 0$ for one value of u, we will generally have $\sigma^0 \neq 0$ for a

[9] In fact, Sachs has pointed out (Ref. 6) that the translation subgroup of the BMS group is uniquely singled out by group theoretic considerations, namely as the only four-parameter normal subgroup of the BMS group.

later value of u (i.e., "goodness" would not be invariant under time translation). The idea of the present paper is that, if we make apparently reasonable physical assumptions as to how the gravitational radiation falls off at $u = -\infty$ (or alternatively at $u = +\infty$), then we can effectively minimize σ^0 at $u = -\infty$ (or alternatively at $u = +\infty$). This will restrict our coordinates to such an extent that only a subgroup of the BMS group remains—which is isomorphic to the improper orthochronous inhomogeneous Lorentz group.

III. SPIN-WEIGHTED FUNCTIONS ON A SPHERE

In order to analyze the structure of σ^0 as a function of the angular coordinates θ, ϕ, it is important first to realize that, although it is given as a scalar, it is really a tensorlike quantity. The directions of minimum and maximum shear are determined by $\arg \sigma^0$. Under rotation of the spacelike vectors Re (m^a), Im (m^a) in their plane given by

$$(m^a)' = e^{i\psi} m^a \tag{3.1}$$

[with $(l^a)' = l^a$, ψ real], we then have

$$(\sigma^0)' = e^{2i\psi}\sigma^0. \tag{3.2}$$

We shall say that σ^0 has spin weight 2. Generally, a quantity η will be said to have *spin weight* s if it transforms as

$$\eta' = e^{si\psi}\eta \tag{3.3}$$

under (3.1). (Here, s is in general integral, but half integral values can also occur.) The quantity σ^0 has, in addition, a conformal weight of -1. That is, under (2.1) and (2.2), if we choose $\alpha = 0$ and examine the σ^0 for $u = 0$ (keeping the m^a vectors fixed), we find[2]

$$(\sigma^0)' = K^{-1}\sigma^0. \tag{3.4}$$

Generally, a quantity η defined on the (θ, ϕ)-sphere has *conformal weight* w if under conformal transformation of the sphere with conformal factor K as in (2.2) (and with fixed m^a vectors) we have

$$\eta' = K^w\eta. \tag{3.5}$$

[K is, in effect, the relativistic Doppler factor[2] $(c + v)^{\frac{1}{2}}(c - v)^{-\frac{1}{2}}$.] For consistency with the coordinate conditions,[1,2,4] this conformal transformation should be accompanied by

$$(l^a)' = Kl^a, \qquad r' = K^{-1}r, \tag{3.6}$$

with $(m^a)' = m^a$, whence (3.4) [cf. (2.4)].

Effectively, the concepts of spin-weight and conformal weight refer to the behavior of functions on the (θ, ϕ)-sphere at infinity only, and do not refer to the remainder of the space–time. Indeed, the concepts will apply to *any* two-dimensional abstract surface, with a Riemannian or conformal structure. Quantities with spin weights correspond to irreducible tensor quantities on the surface. The vectors Re (m^a), Im (m^a) may be regarded as orthogonal tangent vectors (of length $2^{-\frac{1}{2}}$) at each point of the surface. But we shall be concerned, here, only with quantities defined on a *sphere*. If spherical polar coordinates are used, a natural choice[2,4,7] for m^a is to make Re (m^a) and Im (m^a) tangential, respectively, to the curves $\phi = \text{const}$ and $\theta = \text{const}$. Another convenient coordinate system for the sphere is $(\zeta, \bar{\zeta})$ where the complex parameter ζ is related to (θ, ϕ) by

$$\zeta = e^{i\phi}\cot\tfrac{1}{2}\theta. \tag{3.7}$$

In this case, the natural choice for m^a is to make Re (m^a) and Im (m^a) tangential, respectively, to the curves Im $(\zeta) = \text{const}$ and Re $(\zeta) = \text{const}$.

Let η be a quantity defined on the sphere of spin weight s. Define the operator $\eth$, in a particular (θ, ϕ) coordinate system, by

$$\eth\eta = -(\sin\theta)^s\left\{\frac{\partial}{\partial\theta} + \frac{i}{\sin\theta}\frac{\partial}{\partial\phi}\right\}[(\sin\theta)^{-s}\eta]. \tag{3.8}$$

(The operator $\eth$ is effectively a covariant differentiation operator in the surface.) Under a rotation of m^a (but, for the moment, keeping the coordinates fixed) we demand that $\eth\eta$ behave as a quantity of spin weight $s + 1$. From (3.7) and the fact that the m^a vectors are rotated through an angle $-\phi$ in the passage from (θ, ϕ) to $(\zeta, \bar{\zeta})$, we obtain

$$\eth\eta = 2P^{1-s}\,\partial(P^s\eta)/\partial\zeta \tag{3.9}$$

as the definition of $\eth$ in the $(\zeta, \bar{\zeta})$ system, where

$$P = \tfrac{1}{2}(1 + \zeta\bar{\zeta}). \tag{3.10}$$

[The coordinate ζ and the P used here are related to the ζ and P (now called ζ' and P') of Ref. 8 by $\zeta = -\frac{1}{2}\bar{\zeta}'$, $P = \frac{1}{2}\sqrt{2}P'$.]

Now, the first important property of the operator $\eth$ is that it is invariant (with spin weight unity) under change of coordinate system which preserves the sphere metric

$$ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2 = P^{-2}\, d\zeta\, d\bar{\zeta}. \tag{3.11}$$

This is most easily seen in the $(\zeta, \bar{\zeta})$ system and the result for the (θ, ϕ) system will then follow. With ψ as in (3.1), so that $\eta' = e^{si\psi}\eta$, we have

$$e^{-2i\psi} = \frac{\partial\zeta'/\partial\zeta}{\partial\bar{\zeta}'/\partial\bar{\zeta}}, \tag{3.12}$$

where $(\zeta', \bar{\zeta}')$ is the new coordinate system, with new complex tangent vector $m^{a'}$. Since $ds' = ds$, we have

$$\partial\zeta'/\partial\zeta = e^{-i\psi}\, P'/P \tag{3.13}$$

with $\partial\zeta'/\partial\bar{\zeta} = 0$. From this it follows that $\eth'\eta' \equiv 2P'^{1-s}\partial(P'^{s}\eta')/\partial\zeta' = e^{(s+1)i\psi}\,\eth\eta$, as required. [In fact, this argument applies equally to any surface, with metric written $ds^2 = P^{-2}\,d\zeta\,d\bar{\zeta}$, where P need not have the form (3.10).]

The second important property of $\eth$ concerns its behavior under a conformal transformation. Let η have conformal weight w and spin weight s (both integers or both half odd integers) where $w \geq s$. Then,

$$\eth^{w-s+1}\eta \text{ *is a quantity of conformal weight* } s-1, \tag{3.14}$$

and, of course, spin weight $w + 1$. To prove this, we need only consider conformal transformations of the form

$$\zeta' = a\zeta, \tag{3.15}$$

a being a constant, with $(P')^{-2}\,d\zeta'\,d\bar{\zeta}' = K^2P^{-2}\,d\zeta\,d\bar{\zeta}$ so that

$$K = aP/P' = (1 + \zeta\bar{\zeta})/(a^{-1} + a\zeta\bar{\zeta}). \tag{3.16}$$

[The general conformal self-transformation of the sphere, excluding reflections can be built up from transformations of the type (3.15)—corresponding to a Lorentz velocity transformation in the direction of the z axis—and rotations which preserve the metric. These rotations have been already dealt with, since they correspond to metric preserving coordinate transformations.] It is required to prove that, under (3.15), $(\eth^{w-s+1}\eta)' = K^{s-1}\,\eth^{w-s+1}\eta$. [The m^a vectors are unaltered under (3.15).] This can be done by a kind of inductive argument which is somewhat tedious for general values of $w - s$. For small value of $w - s$ it is not hard to verify directly that

$$(P')^{1-w}\frac{\partial}{\partial\zeta'}\left((P')^{2}\frac{\partial}{\partial\zeta'}\left((P')^{3}\cdots\frac{\partial}{\partial\zeta'}\left((P')^{s}K^{w}\eta\right)\cdots\right)\right)$$
$$= K^{s-1}P^{1-w}\frac{\partial}{\partial\zeta}(\cdots(P^{s}\eta)\cdots),$$

where from (3.16) we use the fact that $P^2\,\partial K^{-1}/\partial\zeta$ is a constant. It is a curious fact that, given η with $w \geq s$, there is just the *one* power of $\eth$ given in (3.14) which yields a quantity correctly transforming (i.e., with a "weight") under a conformal transformation. If $w \geq -s$, then $\bar{\eth}^{w+s+1}\eta$ also correctly transforms with conformal weight $-s - 1$. Here $\bar{\eth}$ is defined [in the $(\zeta, \bar{\zeta})$ system] by

$$\bar{\eth}\eta = 2P^{1+s}\,\partial(P^{-s}\eta)/\partial\bar{\zeta}. \tag{3.17}$$

If $w > |s|$, then we can apply both operators to get a quantity with conformal weight $-w - 2$:

$$\bar{\eth}^{w+s+1}\,\eth^{w-s+1}\eta = \eth^{w-s+1}\,\bar{\eth}^{w+s+1}\eta. \tag{3.18}$$

That these two quantities are equal is not immediately obvious since the operators $\eth$ and $\bar{\eth}$ do not in general commute. In fact, we have

$$(\bar{\eth}\,\eth - \eth\,\bar{\eth})\eta = 2s\eta. \tag{3.19}$$

However, it is a consequence of the following considerations that $\eth^{p}\bar{\eth}^{q}\eta = \bar{\eth}^{q}\eth^{p}\eta$ whenever $q - p = 2s$.

The third important property of $\eth$ concerns its effect on spherical harmonics. Let $Y_{l,m}(l = 0, 1, \cdots; m = -l, \cdots, +l)$ be the usual scalar spherical harmonics. Then we can define the *spin s spherical harmonics*[7,8] (for integral spin) as follows:

$${}_sY_{l,m} = \begin{cases} \left[\dfrac{(l-s)!}{(l+s)!}\right]^{\frac{1}{2}}\eth^{s}Y_{l,m} & (0 \leq s \leq l), \\ (-1)^{s}\left[\dfrac{(l+s)!}{(l-s)!}\right]^{\frac{1}{2}}\bar{\eth}^{-s}Y_{l,m} & (-l \leq s \leq 0). \end{cases} \tag{3.20}$$

The ${}_sY_{l,m}$ are not defined for $|s| > l$. In the $(\zeta, \bar{\zeta})$ system, the spin s harmonics take the form

$${}_sY_{l,m} = \frac{a_{l,m}}{[(l-s)!\,(l+s)!]^{\frac{1}{2}}}(1 + \zeta\bar{\zeta})^{-s}$$
$$\times \sum_{p}\zeta^{p}(-\bar{\zeta})^{p+s-m}\binom{l-s}{p}\binom{l+s}{p+s-m}, \tag{3.21}$$

summed over integral values of p, the $a_{l,m}$ being numerical constants whose exact values are inessential for our purposes. [In fact, (3.21) applies also to "spinor harmonics" for which l, m, and s are all half odd integers.] We have ${}_s\bar{Y}_{l,m} = (-1)^{m+s}\,{}_{-s}Y_{l,-m}$ and, for *all* s with $|s| \leq l$,

$$\eth({}_sY_{l,m}) = [(l-s)(l+s+1)]^{\frac{1}{2}}\,{}_{s+1}Y_{l,m},$$
$$\bar{\eth}({}_sY_{l,m}) = -[(l+s)(l-s+1)]^{\frac{1}{2}}\,{}_{s-1}Y_{l,m}. \tag{3.22}$$

In particular, $\eth$ annihilates ${}_sY_{l,m}$ whenever $l = s$ and $\bar{\eth}$ annihilates ${}_sY_{l,m}$ if $l = -s$. We see, furthermore, that $\eth^{p}$ annihilates any quantity of spin weight s which is composed only of harmonics with $l < s + p$.

The ${}_sY_{l,m}$ are eigenfunctions of the operator $\bar{\eth}\eth$ for each spin weight s:

$$\bar{\eth}\eth({}_sY_{l,m}) = -(l-s)(l+s+1)\,{}_sY_{l,m}. \tag{3.23}$$

(If $s = 0$, $\bar{\eth}\eth$ is essentially the total angular momentum operator.) More generally,

$$\bar{\eth}^p \eth^p({}_sY_{l,m}) = (-1)^p \frac{(l-s)!}{(l-s-p)!} \frac{(l+s+p)!}{(l+s)!} {}_sY_{l,m}. \tag{3.24}$$

For each given s, the ${}_sY_{l,m}$ form a set of orthonormal functions of spin weight s on the sphere:

$$\int {}_sY_{l,m}\, {}_s\bar{Y}_{l',m'}\, dS = \delta_{ll'}\, \delta_{mm'}, \tag{3.25}$$

dS being the surface area element on the sphere. This can be proved (s being an integer) by induction on s using (3.22), the orthonormality of $Y_{l,m}$, and the fact that

$$\int (\eth A)B\, dS = -\int A(\eth B)\, dS \tag{3.26}$$

with (spin weight of A) + (spin weight of B) = -1. [It is immediate from (3.8) that $\int \eth C\, dS = 0$ (where C has spin weight -1) and (3.26) follows.] The orthonormal functions ${}_sY_{l,m}$ are also *complete*, for spin weight s quantities on the sphere. That is, if η has spin weight s and is suitably regular[10] on the sphere, there exist constants $\eta_{l,m}$ such that

$$\eta = \sum_{\substack{l,m \\ l\geq s}} \eta_{l,m}\, {}_sY_{l,m}. \tag{3.27}$$

To see this, without loss of generality choose $s > 0$ and consider $\bar{\eth}^s\eta$. This has spin weight zero, so from the completeness of the $Y_{l,m}$, we have, for some constants $\phi_{l,m}$

$$\bar{\eth}^s\eta = \sum_{\substack{l,m \\ l\geq s}} \phi_{l,m} Y_{l,m}. \tag{3.28}$$

By (3.24), we can write this

$$\bar{\eth}^s\{\eta - \sum_{\substack{l,m \\ l\geq s}} \chi_{l,m}\, \eth^s Y_{l,m}\} = \sum_{\substack{l,m \\ l<s}} \phi_{l,m} Y_{l,m}, \tag{3.29}$$

where

$$\chi_{l,m} = \frac{(l-s)!}{(l+s)!}(-1)^s\phi_{l,m} (l \geq s).$$

Moreover, the right-hand side of (3.29) must vanish since, if we multiply (3.29) by $\bar{Y}_{l',m}(l' < s)$ and integrate over the sphere, we get $\phi_{l',m}$ on the right and $\int \bar{Y}_{l',m}\, \bar{\eth}^s\{\cdots\}\, dS$ on the left. Applying (3.26) and $\eth^s Y_{l',m} = 0 (l' < s)$, we see that this vanishes. Thus, all the ϕ's (with $l' < s$) vanish, so the left-hand side of (3.29) also vanishes. Now, if μ is any suitably regular quantity on the sphere, of spin weight s' where $s' > 0$, then

$$\bar{\eth}\mu = 0 \quad \textit{implies} \quad \mu = 0. \tag{3.30}$$

For, by (3.17), $\bar{\eth}\mu = 0$ implies that $P^{-s'}\mu$ is an analytic function of ζ. Hence $\lim_{|\zeta|\to\infty} |P^{-s'}\mu| > 0$ unless $\mu = 0$. But $s' > 0$, so that $P^{-s'} \to 0$ and μ must be bounded for all ζ. Therefore $\mu = 0$. Repeated application of (3.30) to (3.29) now gives (3.27) as required, with

$$\eta_{l,m} = \{(l+s)!/(l-s)!\}^{\frac{1}{2}}\chi_{l,m}.$$

Another way of expressing this result is: given any suitably regular η on the sphere,[10] of (integral) spin weight $s > 0$, there exists ξ of spin weight zero for which

$$\eta = \eth^s\xi. \tag{3.31}$$

The relevance to this paper of the foregoing analysis lies essentially in this result.[11] (In the above, $\xi = \sum \chi_{l,m}Y_{l,m}$.) One reason this will be of interest to us here is that it enables us to define the "electric" and "magnetic" parts of η, denoted, respectively, by η_e and η_m, where

$$\eta_e = \eth^s \operatorname{Re}(\xi), \qquad \eta_m = \eth^s \operatorname{Im}(\xi). \tag{3.32}$$

[If, instead, $s < 0$, then we would use $\bar{\eth}$ in place of $\eth$ and $-s$ in place of s in (3.31), (3.32).] We would get the same η_e and η_m if a different choice of ξ were made in (3.31), since the arbitrariness in ξ lies only in harmonics $Y_{l,m}$ with $l < s$. We can characterize η_e and η_m by the fact that

$$\eta = \eta_e + \eta_m \tag{3.33}$$

with

$$\bar{\eth}^s\eta_e = \eth^s\bar{\eta}_e, \qquad \bar{\eth}^s\eta_m = -\eth^s\bar{\eta}_m. \tag{3.34}$$

The invariance properties of $\eth$ imply that the concepts of "magnetic" and "electric" parts of η are suitably invariant under rotations of the sphere or of the m^α vectors. Furthermore if (and *only* if) η has conformal weight -1, the splitting of η into its "electric" and "magnetic" parts is also invariant under *conformal* transformation of the sphere [cf. (3.14)].

Here we shall apply these concepts to σ^0. The spin weight is 2 and the conformal weight is -1, so on each hypersurface u = const, we have a splitting $\sigma^0 = \sigma^0_e + \sigma^0_m$, which is invariant under conformal transformations of the (θ, ϕ)-sphere, i.e., under the "Lorentz" transformations (2.1) which leave the hypersurface invariant. The σ^0 and its u-derivatives describe properties of the gravitational

[10] We use the phrase "suitably regular" in the sense that η is a smooth function of ζ and $\bar{\zeta}$ in the $(\zeta, \bar{\zeta})$ system and that if we transform coordinates by, say, $\zeta' = \zeta^{-1}$, then η' is a smooth function of ζ' and $\bar{\zeta}'$. This deals with the point $\zeta = \infty$ in the $(\zeta, \bar{\zeta})$ system.

[11] This result, in the case $\eta = \sigma^0$, is considered by D. Lamb, J. Math. Phys. 7, 458 (1966).

radiation field. By analogy with the linear theory, we may suppose μ_0^0 to be associated with "electric"-type radiation (e.g., arising from changes in the mass quadrupole) and σ_m^0 to be associated with "magnetic"-type radiation (e.g., arising from changes in the angular momentum quadrupole). The analogy with electrodynamics arising here is the reason for the use of the terms "electric" and "magnetic" in (3.32). In the next section, we consider these matters more explicitly.

IV. PHYSICAL ASSUMPTIONS

We wish to impose physical restrictions on the systems under consideration, which have sufficient generality to include scattering problems in which sources can come in from and go out to spatial infinity with initial and final velocities less than c. In particular, the sources might all be confined to a bounded region of space. We shall require that the angular momentum, with respect to some origin, of each component part of the source remains bounded for all time (where it is assumed that there are a finite number of "parts" to the source). Generally, the physical restrictions that we impose are intended to be such that, $u \to -\infty$ (or alternatively as $u \to +\infty$), σ^0 behaves as

$$\sigma^0(u, \theta, \phi) \to S(\theta, \phi), \tag{4.1}$$

with S independent of u. If the analogy with the linear theory is to be trusted, we would normally expect S to have no "magnetic" part. This would add some force to the arguments but it is not actually essential. On the basis of (4.1) alone, we shall be able to extract the inhomogeneous Lorentz group.

In order to investigate what is entailed physically by the requirement that (4.1) hold, we must resort to analogies with the linear theory. The full theory is, unfortunately, insufficiently developed as yet to enable us to infer the exact relation between the motion of the sources and asymptotic quantities such as σ^0. In the linear theory, we may take $u =$ const to be "good" cones in Minkowski space. Since the shear of these cones now vanishes, the linearized σ^0, denoted σ^0_{lin}, describes a first-order deviation from "goodness" of these cones. We can then obtain σ^0_{lin} from the multipole moments $C_{l,m}(u)$ of the source distribution by[6,7]

$$\sigma^0_{\text{lin}} = \eth^2\left\{\sum_{l,m} \beta_{l,m} \frac{d^l}{du^l}\,(C_{l,m}Y_{l,m}) + f\right\}, \tag{4.2}$$

where $f = f(\theta, \phi)$ is real, independent of u, and corresponds to the gauge freedom in σ^0, and where the $\beta_{l,m}$ are some inessential numerics. (Both $\beta_{l,m}$ and $C_{l,m}$ have spin weight zero.)

The type of behavior for the multipole moments that we may assume as reasonable is

$$C^e_{l,m} \sim u^l, \tag{4.3}$$

$$C^m_{l,m} \sim u^{l-1}, \tag{4.4}$$

where $C^e_{l,m}$ and $C^m_{l,m}$ correspond, respectively, to the "electric" and "magnetic" parts of the multipole moments, so that $C^e_{l,m}Y_{l,m}$ and $C^m_{l,m}Y_{l,m}$ are the real and imaginary parts of $C_{l,m}Y_{l,m}$. For example, for $l = 2$, the $C^e_{2,m}$ and $C^m_{2,m}$ are, respectively, the five components of the mass quadrupole moment and spin (or angular momentum) quadrupole moment, defined so that $C^e_{2,m} \sim \sum MR^2$ and $C^m_{2,m} \sim \sum LR$, where M, L, and R are the mass, angular momentum, and distance from the origin of each component of the source. If each component is moving with uniform limiting velocity and each angular momentum is bounded, we have $R \sim u$, $L \sim 1$, whence $C^e_{2,m} \sim u^2$ and $C^m_{2,m} \sim u$. The corresponding argument for higher moments leads to (4.3) and (4.4).

By substituting (4.3) and (4.4) into (4.2), a reasonable guess for σ^0_{lin} results:

$$\sigma^0_{\text{lin}} = \eth^2\{A + iB + f\}, \tag{4.5}$$

where

$$A = \sum_{l,m} \beta_{l,m} \frac{d^l(C^e_{l,m}Y_{l,m})}{du^l}$$

is real and tends to a finite limit as $u \to -\infty$ (or as $u \to +\infty$), and where

$$B = \sum_{l,m} \beta_{l,m} \frac{d^l(C^m_{l,m}Y_{l,m})}{du^l}$$

is real and tends to zero as $u \to -\infty$ (or $u \to +\infty$). Of course, (4.5) cannot be inferred rigorously from (4.3) and (4.4) since it is not permissible simply to differentiate order of magnitudes in this way. But if the behavior (4.3) and (4.4) is sufficiently "smooth" asymptotically, then the deduction is valid. This would rule out oscillatory behavior for the sources[12] in the distant past (or future), since such behavior would certainly contradict (4.5). We assume, here, that the components of the system have no accelerations in the limit $u \to -\infty$ (or $u \to +\infty$) which would lead to a violation of (4.5).

[12] An interesting case that would be excluded by our assumptions is that of two masses revolving about one another in ever decreasing circles as they lose energy by gravitational radiation. Here the angular momentum is unbounded as $u \to -\infty$ and apparently $\sigma^0 \sim u^{3/8}$.

We can rewrite (4.5) as

$$\sigma^0_{\text{lin}} \to S_e(\theta, \phi) \tag{4.6}$$

when $u \to -\infty$ (or $u \to +\infty$), S_e being purely "electric" and independent of u. Thus, it seems not unreasonable to infer that, under similar physical assumptions in the full theory, (4.1) will hold [with $S(\theta, \phi)$ purely electric].

It may be felt that the analogy with the linear theory would be a little more trustworthy, however, if given in terms of gauge independent quantities such as the radiation field $\partial^2\sigma^0/\partial u^2$, rather than σ^0. If we assume $\partial^2\sigma^0/\partial u^2$ behaves, in the full theory, in the same way as a $\partial^2\sigma^0_{\text{lin}}/\partial u^2$ consistent with (4.6), then the only difference in the behavior of σ^0 that could arise would lie in the constants of integration. Moreover, the constant of integration involved in the passage from $\partial^2\sigma^0/\partial u^2$ to $\partial\sigma^0/\partial u$ must be zero. This follows from a requirement that the total energy radiated away, as measured according to the Bondi–Sachs formula[1,2]

$$\iiint \left|\frac{\partial\sigma^0}{\partial u}\right|^2 du\, d\theta \sin\theta\, d\phi,$$

be finite. In fact, this finiteness requirement for the gravitationally radiated energy greatly restricts the behavior of $\partial\sigma^0/\partial u$ as $u \to \pm\infty$ in any case in the full theory, in a *rigorous* way (e.g., any simple oscillatory behavior would be ruled out). However, this is not sufficient for our purposes, since the radiated energy is finite if $\sigma^0 \sim u^{\frac{1}{2}-\epsilon}(0 < \epsilon < \frac{1}{2})$ but this would violate (4.1). The remaining constant of integration arises in the passage from $\partial\sigma^0/\partial u$ to σ^0, and this would be the S of (4.1). In the linear theory the $S = S_e$ is purely "electric" and this may be reasonable in the full theory also. But, in any case, the possibility of S having a "magnetic" part can also be treated here.

We now show that, under the assumptions stated above, it is possible to introduce coordinate conditions such that $S_e(\theta, \phi) = 0$. This will be irrespective of whether the "magnetic" part S_m, of $S(=S_e + S_m)$, vanishes. Sachs[2] has shown that, under a BMS transformation (2.1), σ^0 transforms [cf. also (3.2), (3.4)] as

$$\sigma^{0\prime}(u, \theta, \phi) = K^{-1} e^{2i\psi}(\sigma^0(u, \theta, \phi) - \tfrac{1}{2}\eth^2\alpha) \tag{4.7}$$

(α having spin weight zero). It should be pointed out that $\sigma^{0\prime}(u, \theta, \phi)$ refers to the asymptotic shear of the hypersurfaces $u' = \text{const}$ of the *transformed* coordinate system, although evaluated at the (u, θ, ϕ) of the *original* coordinate system. The complete transformation to $\sigma^{0\prime}(u', \theta', \phi')$ is, of course, more complicated. Applied to (4.1), (4.7) gives

$$S'_e(\theta, \phi) = K^{-1} e^{2i\psi}(S_e(\theta, \phi) - \tfrac{1}{2}\eth^2\alpha), \tag{4.8}$$

$$S'_m(\theta, \phi) = K^{-1} e^{2i\psi} S_m(\theta, \phi), \tag{4.9}$$

since α is real. Since by (3.32) we can write $S_e(\theta, \phi) = \eth^2 G$ for some real G and since α can be chosen arbitrarily on the (θ, ϕ)-sphere, it follows that a BMS transformation exists (e.g., $\alpha = 2G$) for which $S'_e(\theta, \phi) = 0$, as required.

We thus may adopt as our coordinate condition the requirement that $S_e(\theta, \phi) = 0$. This can be done for the S_e defined *either* at $u = -\infty$ *or* at $u = +\infty$. For definiteness we could choose the condition defined at $u = -\infty$. (If $S_m = 0$, this means that in the limit $u \to -\infty$ the hypersurfaces $u = \text{const}$ become nearer and nearer to being asymptotically like the "good" cones of Minkowski space.) The BMS transformations which preserve the coordinate condition $S_e = 0$ are now those for which $\eth^2\alpha = 0$ [cf. (4.8)]. Thus α has the form (2.3) and the allowed supertranslations are simply the translations. The "Lorentz rotations" given by $\alpha = 0$ clearly do not destroy the coordinate condition $S_e = 0$. [The fact that S has conformal weight -1 is essential here, since the splitting $S = S_e + S_m$ must be invariant under conformal transformation of the (θ, ϕ)-sphere, $\eth^2$ being also suitably conformally invariant here.] The transformations (2.1) with α as in (2.3) have the same form as inhomogeneous Lorentz transformations in Minkowski space. Thus, the group of coordinate transformations which preserve all the usual coordinate conditions in asymptotic gravitational radiation theory *as well* as keeping $S_e = 0$ at $u = -\infty$ (or else at $u = +\infty$) is isomorphic to the improper orthochronous inhomogeneous Lorentz group.

V. DISCUSSION

There are several inconclusive aspects of this work (even apart from the heuristic nature of the argument concerning the physical assumptions) which should be mentioned. In particular, the ambiguity as to where the coordinate condition $S_e = 0$ is to be imposed is disturbing. If the condition is imposed at $u = -\infty$, there seems no reason at all to believe that the condition would then automatically hold also at $u = +\infty$ (although it is just conceivable that "coherence" relations between the radiation-field data at infinity, of the type considered by Friedlander,[13] might link the con-

[13] F. G. Friedlander, Proc. Roy. Soc. (London) **A279**, 386 (1964).

dition $S_0 = 0$ at $u = -\infty$ with that at $u = +\infty$). In fact, the situation is worse than this since all the above arguments could be repeated using an advanced time parameter v in place of the retarded time u. We would then have four different alternative ways of choosing coordinate conditions, namely $S_0 = 0$ at $u = \pm\infty$, or the corresponding $S_0 = 0$ at $v = \pm\infty$, any one of which would lead to the inhomogeneous Lorentz group, but the different choices would *apparently* be quite unrelated.

It would be interesting to know whether the conditions at $u = -\infty$ are in any way related to those at $v = +\infty$. In the conformal approach to asymptotic analysis,[5] $v = -\infty$, $u = -\infty$, and $u = +\infty$ appear as distinct "points" I^-, I^0, and I^+, respectively, but $v = +\infty$ represents the *same* "points" I^0 as $u = -\infty$. Another question of interest is the relation of the present work to that of the Bergmann,[14] who concludes that if one examines asymptotically flat space–times by proceeding to infinity in *spacelike* directions, then one cannot extract the inhomogeneous Lorentz group as an asymptotic symmetry group. From the conformal point of view, this is again concerned with the "point" I^0, but, since I^0 is approached quite differently in the two analyses, there appears to be no essential conflict here.

A further unanswered question concerns the relevance of the present work to some recent suggestions[6,15,16] that possibly the BMS group can play a role in elementary-particle physics. It is not at all clear that considerations of this paper affect this possibility. The arguments here require consideration of infinite times, while the BMS group emerges if only finite (retarded) time intervals are considered. For such finite time intervals, the BMS group must still be regarded as the relevant asymptotic symmetry group for situations involving gravitational radiation in asymptotically flat space–time.

ACKNOWLEDGMENT

One of the authors, E. Newman, would like to thank Professor H. Bondi and Kings College, London, for their hospitality while this work was being completed. We would like also to thank Dr. A. Janis for having caught a serious error in the original manuscript.

[14] P. G. Bergmann, Phys. Rev. **124**, 274 (1961).

[15] E. Newman, Nature **206**, 811 (1965).

[16] A. Komar, Phys. Rev. Letters **15**, 76 (1961).

27
SOME NEW GRAVITATIONALLY CONSERVED QUANTITIES

Gravity Research Foundation Essay Prize, 1966.
With Ezra T. Newman

We obtained second prize for this Gravity Research Foundation Essay; see also Chapter 23.

Some New Gravitationally Conserved Quantities

Ezra T. Newman

Physics Department, University of Pittsburgh

Pittsburgh, Pennsylvania

and

Roger Penrose

Mathematics Department, Birkbeck College, University of London

London, England

ABSTRACT

It has been commonly believed that there were no absolutely conserved quantities in the Einstein theory of gravity and that only charge was absolutely conserved in the Einstein-Maxwell theory. It is the purpose of the present note to state that this is wrong and to describe ten recently discovered absolutely conserved quantities in the Einstein theory and six conserved quantities in the Einstein-Maxwell theory. These quantities, though they have a trivial meaning in the linear theory of gravity, appear to play a fundamental role in the dynamics of the non-linear gravitational field.

SOME NEW GRAVITATIONALLY CONSERVED QUANTITIES

We shall describe a set of ten recently discovered conservation laws[1] in the Einstein theory of gravitation whose existence had been previously unsuspected. Although these conserved quantities differ fundamentally from certain counterparts in the linear theory of gravity, (which are essentially trivial) the linear theory plays an essential role here in their interpretation. We also parallel the description of the full and linear theory with an analogous description of Einstein-Maxwell theory where there are six new conserved quantities.

As a preliminary there are two mathematical points to be mentioned. First we define two sets of complete functions on the sphere (different from the spherical harmonics) as follows;

$$ {}^{1}Y_{\ell m} = -\sqrt{\frac{(\ell-1)!}{(\ell+1)!}}\left\{\frac{\partial}{\partial\theta} + \frac{i}{\sin\theta}\frac{\partial}{\partial\phi}\right\} Y_{\ell m}(\theta,\phi), \quad \ell \geqslant 1, $$

$$ {}^{2}Y_{\ell m} = \sqrt{\frac{(\ell-2)!}{(\ell+2)!}}\,\sin\theta\left\{\frac{\partial}{\partial\theta} + \frac{i}{\sin\theta}\frac{\partial}{\partial\phi}\right\}\frac{1}{\sin\theta}\left\{\frac{\partial}{\partial\theta} + \frac{i}{\sin\theta}\frac{\partial}{\partial\phi}\right\} Y_{\ell m}, \quad \ell \geqslant 2. $$

Higher type functions, denoted ${}_{s}Y_{\ell m}$, which are particularly convenient for use with spin s fields, can also be defined.

The second point concerns the formalism used. At least in the flat-space Maxwell and linearized gravitational theory one can, instead of Minkowskian coordinates, use null polar coordinates, polar coordinates θ,ϕ,r and the retarded time $u = t - r$. A generalization of this coordinate system can be introduced in the full theory when the space becomes asymptotically Minkowskian. The hypersurfaces $u = \text{const.}$ remaining null and outgoing.

When investigating radiative solutions of Maxwell's equations it is often simpler to work with the following complex quantities than with $\vec{E}$ and $\vec{B}$;

$$\phi_o = (\vec{E} + i\vec{B})\cdot\vec{m}, \quad \phi_1 = (\vec{E} + i\vec{B})\cdot\vec{c}, \quad \phi_2 = (\vec{E} + i\vec{B})\cdot\bar{\vec{m}} .$$

The vector $\vec{m}$ and its complex conjugate $\bar{\vec{m}}$ are defined from an orthonormal triad $\vec{a}$, $\vec{b}$ and $\vec{c}$ (defined at each point of space) by $\vec{m} = \frac{1}{\sqrt{2}}(\vec{a} + i\vec{b})$. $\vec{a}$ and $\vec{b}$ are tangent vectors to the polar coordinate spheres and $\vec{c}$ is along the radial direction.

In the linear and full theory of gravitation, the quantity analogous to $F_{\mu\nu}$ is the Weyl tensor, $C_{\alpha\beta\gamma\delta}$. Just as $F_{\mu\nu}$ corresponds to $(\vec{E}, \vec{B})$, the $C_{\alpha\beta\gamma\delta}$ corresponds to two traceless symmetric three dimensional tensors. V_{ij} and W_{ij}. We define five complex quantities analogous to the above φ's;

$$\psi_o = (V_{ij} + iW_{ij})m_i m_j, \qquad \psi_1 = (V_{ij} + iW_{ij})c_i m_j$$

$$\psi_2 = (V_{ij} + iW_{ij})(c_i c_j + m_i \bar{m}_j), \quad \psi_3 = (V_{ij} + iW_{ij})\bar{m}_i c_j$$

$$\psi_4 = (V_{ij} + iW_{ij})\bar{m}_j \bar{m}_j \quad .$$

The differential equations satisfied by the ϕ's and ψ's can be integrated asymptotically without great difficulty. The solutions for all outgoing waves and those incoming waves which concern us have the asymptotic form;

$$\phi_o = \frac{\phi_o^\circ}{r^3} + \frac{\phi_o'}{r^4} + O(r^{-5}), \qquad \phi_1 = \frac{\phi_1^\circ}{r^2} + O(r^{-3}) ,$$

$$\phi_2 = \frac{\phi_2^\circ}{r} + O(r^{-2}) ,$$

$$\Psi_o = \frac{\Psi_o^\circ}{r^5} + \frac{\Psi_o^1}{r^6} + O(r^{-7}), \quad \Psi_1 = \frac{\Psi_1^\circ}{r^4} + O(r^{-5})$$

$$\Psi_2 = \frac{\Psi_2^\circ}{r^3} + O(r^{-4}), \qquad \Psi_3 = \frac{\Psi_3^\circ}{r^2} + O(r^{-3}) \quad ,$$

$$\Psi_4 = \frac{\Psi_4^\circ}{r} + O(r^{-2}) \quad .$$

This general asymptotic behavior holds rigorously for the full theory and for the Einstein-Maxwell theory as well as for the linear theories.

The usual conserved quantities in the linear theories, namely charge, mass, and linear and angular momentum, can be given as surface integrals at fixed u of ϕ_1°, Ψ_2° and Ψ_1°, with appropriate ${}_sY_{\ell m}$ as weighting factors.

The new conserved quantities can be similarly defined; for the Maxwell theory they are (in complex form)

$$F_m = \int \phi_o^1 \; {}_1\bar{Y}_{1m} \, d\Omega, \qquad m = -1, 0, 1,$$

and for the linear and full theory of gravitation

$$G_m = \int \Psi_o^1 \; {}_2\bar{Y}_{2m} \, d\Omega, \quad m = -2, -1, 0, 1, 2 \quad .$$

In the full Einstein-Maxwell theory, the three F_m are still strictly conserved but the G_m (as defined above) are no longer conserved but can be "carried away" by electromagnetic radiation. If the total charge of the system vanishes, then an electromagnetic contribution R_m to the G_m can be defined such that the $G_m + R_m$ is conserved. The G_m constitute a D(2,0) representation of the homogeneous Lorentz group; they are unchanged by translations and by what is known as supertranslations.

To understand the F_m and G_m it is useful to first investigate their meaning in the linear theories. It is easy to show (in this case) that they both vanish for pure retarded fields. In addition they vanish for all advanced fields except for the dipole field in Maxwell theory and

the quadrupole field in the linear gravitational case. In fact if the time profiles (of advanced time, $v = t + r$) of the advanced fields behave as $\frac{\alpha_m}{v} + O(v^{-2})$ for large v, the F_m and G_m are proportional to α_m. In other words, in the linear theories, the F_m and G_m measure the existence (in the asymptotic null future) of incoming dipole and quadrupole waves which are dying out as $1/v$.

Initially we believed that this trivial interpretation of the G_m extended to the full theory. However it was here that the non-linearity of the full theory presented an essentially new situation.

Consider the static or stationary vacuum solutions of the full theory, where incoming (or outgoing) radiation is absent according to any reasonable definition - nevertheless, the G_m do <u>not</u> vanish. They can be expressed in terms of the monopole, dipole and quadrupole moments, m, p_i and Q_{ij}, by the standard[2] relation $G_m \longleftrightarrow K_{ij}$, where

$$K_{ij} = m\, Q_{ij} - 3p_i p_j + p^2 \delta_{ij} \quad .$$

The imaginary parts of p_i and Q_{ij}, are the spin dipole and quadrupole moments respectively. K_{ij} is a trace free tensor, independent of the origin with respect to which the Q_{ij} and p_i are calculated.

This result for the stationary solutions leads to a remarkable conclusion. Suppose an initially stationary asymmetric body with $K_{ij} \neq 0$ becomes spherically symmetric at a finite time later. Since the Q_{ij} and p_i must then vanish, K_{ij} must become zero but the G_m must still correspond to the <u>original</u> values of K_{ij}. This leads to a time dependent solution containing terms very similar to the incoming waves of the linear theory, but now presumably back-scattered waves. It will not become stationary <u>in any finite time</u>. To make it stationary the system would have to develop

moments such that the new K_{ij} would equal the old. We see that the G_m thus make certain classical stationary states inaccessible in a finite time from other states.

It appears therefore as if the G_m and F_m represent in addition to information about incoming waves, information about the source; the relative importance of each meaning depending on details in each case.

PHYSICAL IMPLICATIONS

We now have a set of quantities which are absolutely conserved in asymptotically flat space-times satisfying the appropriate Einstein equations and which can be measured by an examination of the asymptotic fields. The only previously known quantity of this kind was electric charge. In view of the importance to physics of the concept of charge, it seems not unreasonable to expect that there may be significant physical consequence of G_m and F_m conservation, also.

Indeed, the conservation of G_m leads immediately to one interesting conclusion concerning gravitational radiation. For the first time we have a rigorous argument which shows that gravitatonal waves emanating from an isolated system must in general undergo back-scattering. Thus a transition from a stationary state to another state with a different K_{ij} can never be achieved with clean-cut waves; the second state never achieves exact stationarity showing that there must be a residual gravitational disturbance within the light cone of the wave. (The hope had sometimes been expressed that gravitational waves might in some sense strictly obey Huygen's principle, i.e. be without tails.

When it comes to examining the role and meaning of the new conservation laws in actual physical processes, the picture unfortunately becomes more difficult and obscure. We may speculate on the role of G_m in the study of massive objects and the question of gravitational collapse, since there gravitational forces become significant in comparison with other forces. We are familiar with the idea of mass-energy conservation being of great importance in virtually any physical situation. Although conservation of mass-energy in general relativity is not so clear-cut as it is in special relativity, we can say at least the <u>total</u> mass-energy

of an isolated system will be conserved provided we take into account a (positive) contribution to the energy which resides in the emitted gravitational waves. Now, in effect, this energy of the gravitational waves is lost to the system owing to the weakness of the gravitational coupling. During an asymmetrical collapse, large reductions in the mass of the material system may occur quite rapidly (if calculations based on the linear theory are to be believed.) Thus, mass-energy conservation will not effectively be operative at this stage. On the other hand, G_m conservation may be expected to be relatively more important since G_m is not carried away by the waves. It is possible that the behavior of a collapsing system may be severely restricted by the G_m conservation. A great deal more work needs to be done to see if the conserved quantities do lead to important physical effects in phenomenon of collapse.

An alternate possibility is that the G_m and F_m might play a role in phenomena on an extremely small scale. One might speculate as to whether the conserved quantities lead to the existence of conserved currents, (as does charge conservation) which might perhaps, act as the source of undercovered new fields. In the absence of any larger theory, such as a transformation theory which might derive the conservation laws by an application of Noether's theorem, we must at present leave the matter here.

REFERENCES

1. E. Newman and R. Penrose, Phys. Rev. Letters 15, 231 (1965).
2. J. D. Jackson, Classical Electrodynamics, John Wiley & Sons, New York, page 99.

28
AN ANALYSIS OF THE STRUCTURE OF SPACE-TIME

Essay, Awarded the Adams Prize 1965–66. Duplicated by Princeton University, 1967

This essay was awarded the Adams Prize of Cambridge University 1965–66. A major part of this essay is reprinted in the 'Battelle Lectures' that constitute Chapter 36, in Volume 2 of these *Collected Works*, and that provided the official write-up required for the Prize. Stephen Hawking was awarded an auxiliary Adams Prize at the same time, which has perhaps been a source of some confusion.

AN ANALYSIS OF THE STRUCTURE OF SPACE-TIME *

by

Roger Penrose

Palmer Physical Laboratory
Princeton University
Princeton, New Jersey

and

Department of Mathematics
Birkbeck College
University of London
London, England

Work partly supported by the U. S. Air Force Office of Research, Air Research and Development Command under contract number AF49(245)-6013.

Abstract i

I Theories of Space-time 1

II Mathematics for Curved Space-time 17

III Two-Component Spinor Formalism 31

IV Exact Sets of Fields 40

V Initial Data on a Null Cone 50

VI The Spin Coefficient Method 61

VII The Conformal Technique 81

VIII Collapse and Cosmology 94

Appendix 123

Footnotes and References 128

Figures

1 36

2,3 83a

4, 5, 6, 7 84a

8, 9 90a

10, 11, 12, 13 92a

14, 15, 16, 17 97a

18 111

19, 20 120a

A1. 124

A2, A3, A4, A5. 125

A6, A7, A8, A9. 126

Abstract

This essay is concerned with techniques for analysing the structure of space-time and with certain applications. The motivation, in the first instance, for the introduction of the particular techniques used here, springs from a belief that perhaps ultimately the differentiable manifold picture of space-time may have to be abandoned; or, at least, that the null cone and "causality" structure of space-time should in some way be regarded as essentially more fundamental than its continuity-differentiability structure. The idea, then, is to reformulate the relativity theory in certain ways, so that the emphasis is transferred from coordinate systems and from the methods of general differential geometry to certain concepts which are tied to the existence of the null cones and even to the particular dimensionality and signature of our space-time. Thus the methods which emerge have to be more tailored to relativity physics than are those techniques which may be regarded as presently "conventional".

There are three main ideas introduced here. In the first instance, van der Waerden two-component spinors are used for the description of space-time geometry, the essential point being that spinors represent a breakdown of tensors into more elementary parts. With spinors, it is null directions which are described most naturally; the particular null cone structure of (1 + 3)-dimensional space-time becomes built into the formalism at the outset. The second idea is, to regard conformal structure as being, for some purposes, more significant than the actual metric structure of space-time. It is the conformal structure which is defined by the null cones, and the conformal attitude leads to a useful means of treating asymptotic questions in relativity. The third idea is to exploit the causality relations in general-relativistic manifolds and to use them to obtain global results of a topological or quasi-topological nature.

The essay consists of eight sections. Section I is concerned with general considerations about the nature of space-time, about how the prevailing view of space and time has evolved over the centuries and about how it may possibly evolve in the future. Together with section II it serves as a partial motivation for the remaining sections. In section II, the present "differentiable manifold" view of space-time is analyzed. A reformulation of the topology concept, in terms of the more immediately physical idea of causality is presented. The basic physical motivations behind the adoption of a spinor formalism are also given. In section III the details of the spinor method are introduced with an emphasis on a coordinate independent approach. In section IV general systems of physical fields are discussed in the spinor formalism. Gravitation is described by a spinor which represents the Weyl curvature and is treated on a similar footing to other fields. The concept of an exact set of fields is defined and it is shown in section V that this concept is a very appropriate one for the discussion of a constraint-free initial value problem on a null cone. Degrees of freedom for the fields may be obtained simply by counting the spinors which make up the exact set. In section VI the spinor method is developed in a way which is useful for explicit calculations. The method ("spin coefficients") is used to treat Einstein-Maxwell asymptotically flat space-times. With the help of a certain angular differential operator ($\eth$), the Bondi-Sachs energy-momentum law is proved and the time-translation invariance of the Newman-Penrose absolutely conserved quantities is obtained. Section V introduces the technique whereby conformal transformations may be used to study asymptotic questions. The definition of asymptotic simplicity is given, which leads to a coordinate-free definition of asymptotic flatness for vacuum fields. The isotropic

relativistic vanishing pressure cosmologies are briefly analyzed according to this technique, particularly in connection with their "visual horizons". In section VIII is a discussion of gravitational collapse, with particular reference to causality properties and to the occurrence of singularities. The concept of a "semi-spacelike boundary" is introduced. The structure of such boundaries is analyzed and is used to prove, in particular, a very general theorem which tells us that present-day physical laws do not enable us in principle to calculate the future of a system in gravitational collapse. A possible type of implication for cosmology is suggested.

The main ideas of sections (III,) IV, V, VII are original with this author, although they have appeared in a different form earlier. Section VII is essentially taken from two articles by the author (see refs. 46, 66). Sections III, IV, V are extended versions of an unpublished but widely circulated article (see ref. 44). Section VI consists essentially of joint work by the author and E.T. Newman (and A. Exton), not all of which is yet published (see refs. 59, 60, 62, 63). The detailed sentiments expressed in sections I and II have not (to my knowledge) appeared in print before in exactly this form. A part (cf. (2.2)-(2.16)) of section II represents joint work with E.H. Kronheimer(ref. 29). Section VIII is essentially original and contains details, not published before, of the author's theorem on gravitational collapse mentioned earlier (cf. ref. 79). As to relative importance, probably the ideas of section VIII would be regarded as presently of the most significance. My own feeling is that while this may be so, the results of this type of approach are useful mainly in a negative or "destructive" context. Sections IV and V, together possibly with VII and the motivations of section I and of the appendix, present, at least, the potentialities of a constructive approach in the development of space-time theory.

I. THEORIES OF SPACE-TIME

It is natural that man should have long pondered on the nature of space and time. For, is it not true that space and time constitute the environment in which all the phenomena of our experience take place? Or can we, perhaps, go further than this and even identify those phenomena with irregularities in the structure of space (or time)? An idea of this kind appears to have been part of the driving force behind Riemann's introduction, in 1854, of his theory of differential geometry[1,2]. With this theory, Riemann introduced a new flexibility into the concept of geometry; a flexibility which in principle, at least, allows for the possibility of a much more intimate relation between space and matter. However, it was not until 1916 that Einstein's description of gravitation in terms of space-time curvature[3] gave some real physical content to this idea. And it is Einstein's general theory which now tells us that the differential geometry of Riemann describes most accurately the structure of space and time.

That some understanding of space and time is of fundamental importance to the understanding of all natural phenomena, there can never have been much doubt. But on the extent of the role played by space-time in nature, and on the exact structure which this space-time is supposed to possess, many different attitudes have been expressed. Indeed, the prevailing view of space-time has undergone many changes since antiquity. This is partly due to advances in physical understanding, but is to at least as great an extent due to developments in mathematical ideas.

Some of the earliest significant deliberations on the nature of space and time are perhaps contained in the famous "paradoxes" of Zeno. The fact that, today, we do not regard these as any longer strictly paradoxical is an indication of the extent to which our ideas of space and time have changed. Zeno's remarks were aimed at ideas which were prevalent at that time. His paradoxes would appear to have been partly instrumental in the later developments which eventually led to the rigorous mathematical concepts of limit and of the real number continuum. One can fairly say that the

prevailing view of space and of time is, today, based on the notion that space and time are continuous and that a space-time continuum can be accurately described in terms of real numbers, constituting, in fact, what we would now call a differentiable manifold[4]. Perhaps it would be more accurate to say that this differentiable manifold is a mathematical model of those aspects of nature which we describe by the terms "space" and "time". But it is a model of very great accuracy, of such accuracy, in fact, that there is no inconsistency known between the model and experimental fact. Indeed, it would be hard to envisage what form an acutal inconsistency could now take, since the term "space-time" (which in post-relativistic physics must replace the individual "space" and "time") has come to refer, perhaps, more to the model itself than to nature, since in nature the meaning of such terms must inevitably remain somewhat vague. Nevertheless, we should not close our minds to the strong possibility that some future theory may be found which describes nature even more accurately and in a more unified way than present theory, but for which the differentiable manifold picture of space-time would be no longer appropriate.

I shall return to this question shortly, but for the moment, let us accept the picture of space-time as a differentiable manifold and examine the geometric structure that is implied for this manifold, by the various relevant physical theories which have prevailed since ancient times. We can detect five major stages in development, which can be conveniently labelled as follows:

(1.1a) Aristotelian space-time

(1.1b) Galilean space-time

(1.1c) Newtonian space-time

(1.1d) Minkowskian space-time

(1.1a) Einsteinian space-time

In each case, the space-time is to be a four-dimensional differentiable manifold but with additional local geometric structure which will incorporate some aspect of physics. The elements (or "points") of a space-time are referred to as events,

an event being something that one pictures as a point in space, but which has only an instantaneous existence. The history of a particle defines a continuous sequence of events in space-time (a "curve" in space-time) which is called the world-line of the particle.

By Aristotelian space-time, I mean the product $E^3 \times E^1$ of ordinary three-dimensional Euclidean space E^3 ("space") with a one-dimensional Euclidean space E^1 ("time"). (By a Euclidean space, I mean a space carrying a Euclidean metric in the normal way and possessing the usual Euclidean group of motions. Thus, in particular, no point of origin is to be singled out in the space.) The product $E^3 \times E^1$ has the structure that it possesses two metrics, namely those naturally induced by the metric in E^3 and by the metric in E^1. Thus, for any two events in $E^3 \times E^1$ there is a unique spatial separation defined and also a unique time difference. This expresses the fact that in Aristotelian dynamics there is a well-defined concept of simultaneity (i.e. when the time difference is zero) while also the state of rest is to be distinguished from all other motions. (The world line of a particle at rest is a set of events any two of which have zero spatial separation.) The group of motions of $E^3 \times E^1$ is transitive, being the direct product of the Euclidean group on E^3 with that on E^1.

Galilean and Newtonian space-times differ from Aristotelian space-time in that for two general events only one separation exists, namely their time-difference. The spatial separation is not defined for two events with a non-zero time-difference, there being no natural "state of rest", which would be required to define a concept of "the same place at different times". On the other hand, when the time-difference vanishes, a unique spatial separation is defined. The set of all events whose time-difference with a given event vanishes has the property that for any two events of the set the time separation vanishes (equivalence relation). I shall call such a set a simultaneity section of the space-time. The spatial metric defined on each simultaneity section is that of E^3. Through each event passes one and only

one simultaneity section, the "time" E^1 being the factor space of the entire space-time by the simultaneity sections.

To distinguish the Galilean space-time from Newtonian space-time, we must impose some further structure. This is obtained from the physical concept of inertial motion. According to the mechanics of Galilei-Newton "a particle which is not acted on by a force moves in a straight line with uniform velocity". With the hindsight afforded us by Einstein's general theory, we can interpret this as "the world line of a particle not acted on by a force is a geodesic in space-time". Once we have specified the geodesics, we shall have completed the definition of the structure of the space-time. To obtain the Galilean space-time, we specify the geodesics as follows: Consider ordinary Euclidean four-space E^4 and single out a maximal set of parallel E^3's. Now, the factor space of E^4 by these E^3's has an obvious E^1 metric structure. Also the geodesics in E^4 are just the straight lines in E^4. We define the structure of Galilean space-time, then, to be isomorphic in respect of its geodesics, with that of E^4, where its simultaneity sections and time structure correspond respectively, to the parallel E^3's and the factor space E^1. (This construction also gives geodesics in the simultaneity sections. These are, of course, defined anyway by the metric structure, but they do not have direct physical interpretation as inertial world-lines). The group of transformations of Galilean space-time is the 10-parameter Galilei group.

To define the geodesics for Newtonian space-time[5] is more difficult. The intention here is to incorporate the whole of Newtonian gravitational theory.[6] Again we borrow an important idea from Einstein's general theory, namely that gravitation is not to be considered as a true force, but as an inertial or "fictional" force along with the acceleration, centrifugal and Coriolis forces. This, of course, is only possible because of the Galilei-Einstein equivalence principle[6a], which asserts that the inertial mass and the passive gravitational mass of a body are equal.[7] We cannot directly use the same idea to eliminate any other of the (known) forces of

nature. Thus, a particle moving under the action of gravity, but which is not acted on by any other force of nature, is considered to be moving inertially. We are, in fact, more or less compelled to take this point of view by the principle of equivalence, because of the geometric attitude adopted. The world lines of the inertial particles now define the geodesics in Newtonian space-time. Unlike the Aristotelian and Galilean cases just considered (and the Minkowskian one to follow), we cannot, of course, give a complete construction for the space-time, there being many different possible Newtonian space-times corresponding to all the different possible inequivalent gravitational fields. The Galilean space-time is, in fact, just one possible example of a Newtonian space-time, namely that corresponding to zero gravitational field, - or, equivalently, for example, to a field of constant gravitational "acceleration" defined throughout space-time. For, as in general relativity, the physically meaningful aspect of a gravitational field is contained only in the "tidal force" which results from a non-uniform gravitational acceleration field. These tidal forces find their complete description in terms of the curvature of the Newtonian space-time. This curvature measures the extent to which the structure afforded by the geodesics differs from that of Galilean space-time. It is not my purpose here to go into the precise definition of the field equations for Newtonian space-time, which relate the geodesics to the distribution of gravitating matter. Suffice it to say that these equations can be obtained by a limiting operation applied to the equations of general relativity, corresponding to the limit $c \to \infty$, where c denotes the velocity of light. For details, the reader is referred, particularly, to the articles of Trautman.[6]

With Minkowskian and Einsteinian space-times there is a radical departure. We no longer have a transitive concept of simultaneity, so that simultaneity sections do not exist. There is no additive concept of time-difference defined between pairs of events (and, in fact, a natural time-difference concept does not even apply at all to all pairs of events). Instead, we have a definition of time interval which

applies to any pair of events connected by an allowable particle world-line but which depends on the choice of this world-line connecting the events. The time interval is supposed to be that measured by an idealized clock having this as its world-line. The fact that the time interval between two events depends on the route through space-time taken by the idealized clock used to measure it, is often expressed popularly in terms of the "paradox of the twins": The twin who stays at home experiences a longer time interval than his brother who returns home after making a long journey at a speed approaching that of light. The twins have thus aged by different amounts when they meet again. This new "relativistic" concept of time[7a] now serves to define the geometry of the space-time completely. It gives rise to a pseudo-Riemannian metric of hyperbolic-normal signature, the (timelike) geodesics of which define the inertial motions of particles.

The relationship between the Minkowskian and Einsteinian space-time mirrors that between the Galilean and Newtonian cases just considered. Thus the Minkowskian space-time can be described exactly and it does not incorporate gravitation. Its geodesic structure, for example, is the same as that of E^4, the set of timelike geodesics corresponding to geodesics in E^4 making an angle of less than 45° with some fixed direction. Its metric structure can also be defined from that E^4 in relation to this direction. Its 10-parameter group of motions is the Poincaré (i.e. inhomogeneous Lorentz) group.

On the other hand, there are many Einsteinian space-times, one corresponding to each allowable gravitational field. We now have field equations which describe how the matter density relates to this gravitational field. And as with the Newtonian case, it will be the tidal gravitational force which has physical significance, resulting, as before, from a gradient of the gravitational acceleration field. And, as before, the tidal gravitational force will be directly interpretable in terms of an important aspect of the geometry, namely the curvature of space-time.

A detailed mathematical treatment of Einsteinian space-times will be given in

subsequent sections. My purpose up until now has been merely to indicate the essential features of the space-time geometries appropriate to the prevalent physical theories of different eras. Of course, the description in terms of space-time geometry was not, before Minkowski, the way these theories were viewed by their authors (presumably). Nor, even today, is it common to think in such terms, even though the Einstein theory, which is the one I have termed "prevalent" today, is an essentially space-time theory. The tendency is still to think in "Galilean", or more usually, "Aristotelian" terms. (It is perhaps not unreasonable to use such "Aristotelian" concepts when describing everyday occurrences on a large, practically rigid body such as the Earth, but the common descriptions of astronautical achievements in terms of "distances travelled" becomes somewhat confusing!) However there can be little doubt that of all views of space and time that have been put forward up to the present day, the Einsteinian view is the most comprehensive, the most profound and also the most accurate.[8]

But what of future theories? We can envisage two types of directions for development. The first would lie in possible modifications of a comparitively "harmless" type, that is, of a kind which would leave intact the basic premises of the "Einsteinian view". In this category would belong the various unified field theories[9] (including those which are "five-dimensional"[9a]), the theory with asymmetric affine connection for describing spin[10] and the theories which admit a scalar component to the gravitational field.[11] An alternative to such possibilitie would be a much more drastic overhauling of basic principles. Conceivably, this could be of a "retrograde" character, such as the reintroduction of a Minkowskian background[12] as the "true space-time" (or for that matter, why not even an Aristotelian background?) in which gravitation might appear, say, as an ordinary field. But this would seem contrary to the way physics progresses. Much more probably, any successful new theory of space-time would have to retain the advantage of the profound physical insights that Einstein's theory has revealed. It is, nevertheless,

possible to imagine a radically new type of such theory arising: a theory of space-time which would not be based, at the outset, on those concepts which grew out of the resolution of Zeno's paradoxes, namely the concepts of continuity and of a differentiable manifold.

But why is it necessary to change Einstein's theory at all? Is it not conceivable that the Einsteinian view represents the ultimate and "correct" representation of space and time? It is, indeed, conceivable, but in my opinion not very likely. I can give two reasons for my beliefs here, but I would certainly not expect the reader to be necessarily convinced by them. Real conviction would presumably only follow the experimental verification of a theory which accords with such beliefs.

The first reason I can give is somewhat indirect and is of a metaphysico-mathematical nature. It concerns the nature of the continuum itself. The second reason will be of a somewhat more direct physical character. Let us first consider the mathematical nature of a one-dimensional continuum (or real number system) in abstractia. Starting from the non-negative integers (which would surely have a direct interpretation in virtually any physical theory), we introduce new concepts in three stages. First, the negative integers are formally defined - so as to give us the additive group of all integers. (It appears that electric charge, for example, is a physical concept accurately described by the additive group of integers - whether or not quarks exist!) At the next stage, formal quotients are defined, with equivalence relations between them, so as to give the field of all fractions. (Although convenient mathematically, the field of fractions does not seem to have any direct realization in physical terms. It serves merely as an intermediate stage in the mathematical development of the continuum.) Finally the real numbers are introduced, being defined in terms of infinite classes of fractions having certain properties. It is at this final stage that conceptual mathematical difficulties arise owing to essentially non-constructive nature of the definition.

Brower, Weyl and others[13] attempted a redefinition of real numbers in more constructive terms, and while such a definition is possible, the resulting system has more complicated properties than one would like to have for real numbers. In fact, the essential desirable feature that one requires, namely that every increasing sequence of real numbers which is bounded above must have a limit, is not satisfied[13a] by a "constructive continuum". (This is just the property required for the resolution of Zeno's paradoxes!.) As has been emphasized, for example, by Schrödinger,[14] it is the long familiarity of the continuum concept which makes it appear comparitively simple, while its detailed infinitesimal structure - and it is here that the non-constructiveness[15] must enter - has very little actually to do with physics. It is a belief in the ultimate simplicity of nature, then, which, is to me an almost compelling reason for rejecting the differentiable manifold concept as an "ultimate" view of space-time.

But even if this is accepted, it tells us very little, in itself, since there is no indication of the level at which we would expect continuity and differentiability to break down. Let us, then, consider a possible physical reason for the breakdown of the differentiable manifold concept. As a classical theory, general relativity is remarkably complete and self-consistent. But we know that classical laws are not adequate for the description of a large portion of physics. So the question arises of the quantization of general relativity. It seems clear that one cannot just ignore the problem by aruging that: while quantum mechanics refers to microscopic phenomena, gravitation is essentially macroscopic. This perhaps, in some rough way, is the reason that quantum mechanics and general relativity can live together in the same universe, without apparently having much bearing on one another, but it cannot be any real answer. Quantum mechanics does not in itself require that it should apply only to bodies of small mass (say $\ll 10^{-5}$ grams), nor does general relativity assert that only large masses (say $\gg 10^{-5}$ grams) can produce a measurable gravitational effect. (In fact quantum mechanical effects on a macroscopic scale are already

observed in such phenomena as superfluidity, superconductivity and the Josephson current.)

It seems that there are two, or possibly three, attitudes available to take on this question. The first is that one should just proceed and quantize general relativity; that, although there may be formidable difficulties in the way of achieving this, these difficulties are all of the "technical" type, i.e. that there is nothing new of principle to be gained by the exercise. One might be tempted to ask: "then why bother?" - but this would be to presuppose that there are no situations in which both quantum and gravitational (curvature) effects are simultaneously important. There are, in fact, at least two places where we might reasonably expect both types of effect to be significant at once: for example, in possibly defining a gravitational "cut-off" to eliminate divergences in quantum field theory[16]; or at the singular states that appear to be predicted by general relativity either as an "end-point" to gravitational collapse or, perhaps, as an initial state in a "big bang" cosmology.[17] (The question of the existence of such singular states will be discussed in some generality in section VIII.)

The second attitude would be that quantum mechanics and general relativity cannot, or at least should not, be forced together into each other's framework by such a "shotgun wedding" technique; that what is required is something more in the line of a "grand synthesis", i.e. a new theory in which general relativity and quantum theory would emerge as different limiting cases, each applicable to its appropriate domain of phenomena, and in which, hopefully, semi-philosophical quantum mechanical questions such as the meaning of an "observation" might be resolved. In fact, this second point of view is the one to which I would, myself, more readily incline. But it is, for the present, possibly something of the lazy man's way out, since it provides the relativist with an excuse for not tackling directly the substantial problems of quantization! Finally, I should mention that I regard this view to be qualitatively different from that apparently expressed by

An Analysis of the Structure of Space-Time

By Roger Penrose

On page 11, last paragraph, the term "geometry" refers just to some local structure (or tensor fields) on the differentiable manifold. It is not intended to be suggested here that "geometry" means "metric" in the usual sense.

Einstein himself, namely, that quantum effects ought to emerge from a correctly chosen unified field theory - which would be necessarily based on the concept of a differentiable manifold.

It is entirely possible that a "grand synthesis" as suggested above, might have as its basis, something other than the differentiable manifold concept. It might, for example, be based on combinatorial principles, rather than on continuity and differentiability. In the absence of such a theory this is, of course, speculation. But what about the more convential "straightforward" quantization of general relativity? Is this not based on the idea of the differentiable manifold concept? In so far as the quantum structure depends essentially on the nature of the classical structure from which it arises, this is of course true. But the detailed structure of a quantized space-time may be very different from that of a manifold. At least, this is what seems to be implied by certain general considerations of the consequences of quantization (although it may depend on the exact method of quantization used).

Suppose we have two states each of which is represented as a differentiable manifold with a certain geometric structure - in short, by a geometry. The customary "covariant" view would be to regard the states as different if and only if the corresponding geometries were not isomorphic. Now, according to quantum theory, if two different states are possible, then they belong to a whole (connected) two-dimensional array of other states[18], which are described as "linear combinations" of the original two states (with complex coefficients). Thus, if all individual states are to be representable as geometries, we must have a way of forming "linear combinations" of geometries. Unless one has a specific one-to-one mapping between the points of any two geometries automatically defined (and this would seem to go against the spirit of covariance) it is difficult to conceive of how this could be done in any natural way, while retaining the picture of a geometry as a (four-dimensional) manifold with structure. This argument is not conclusive, but the

difficulty is brought out particularly if we allow our manifolds to have differing topologies. Wheeler[19] has suggested that the quantum fluctuations of metric should produce alterations of topology at distances of the order of 10^{-33} cm. This results in a foam-like structure for space-time, since all possible such alterations in topology have to contribute simultaneously in the total picture. The resulting structure thus seems very different from a manifold on the small scale.

Let us, then, briefly list some general alternatives for the structure of space-time in the small:

(1.2a) smooth manifold,

(1.2b) foam structure,

(1.2c) discrete space-time,

(1.2d) no space-time.

Possibilities (1.2a) and (1.2b) have already been mentioned. Included under (1.2c) would, for example, be the lattice space-time of Schild[20] and the theory of Ahmavaara[21] which attempts to replace the real number continuum by a very large but finite field. This second idea, while ingenious, strikes me as too artificial to be really promising. Schild, in his example, pointed out that one could "approximate" the Lorentz group by a certain discrete subgroup of transformations, when a flat Minkowskian lattice of space-time points was used. In this, he attempted to overcome, in part, one of the main problems encountered with the point of view (1.2c), namely that of Lorentz (or Poincaré) covariance for any kind of discrete structure. Nevertheless this problem remains in Schild's example[21a]. (In particular, there is a minimum velocity of $\frac{1}{2}\,3^{-\frac{1}{2}}$ of that of light.) To mention just one other idea in this line, Snyder[21b] considered, non-commuting position operators, so that a quantized space with a discrete spectrum arose. However, in Snyder's model, the time spectrum remained continuous. In attempts at a theory in accordance with (1.2c), it is normally supposed that the discreteness in structure should emerge somewhere at around 10^{-13} cm, this being essentially the radius of an elementary particle. Such a

theory is thus much more radical than the point of view (1.2b), since Wheeler's foam structure occurs at some twenty orders of magnitude smaller.

The point of view of (1.2d) is one expressed by Chew[22]. According to this view, a space-time description of phenomena at dimensions of the order of 10^{-13} cm, or perhaps even at say 10^{-11} cm, is no longer appropriate, so that space-time does not really "exist" at such dimensions. The essential argument is that precise position measurements are impossible in principle to such accuracy because the high momentum required by the uncertainty principle would result in the creation of extra pions or electrons, thus completely altering the situation. Chew regards the momentum space, instead, as being fundamental and attempts to obtain all the required physics from the analyticity of the S-matrix. The argument has some force, particularly in the extent to which experimental facts are incorporated. However, it is difficult to fit in with general relativity as it stands, since in Einstein's theory it is the reverse situation that momentum, rather than position, cannot be defined precisely.

From the standpoint of my own ideas, Chew's view could not be completely satisfactory, in any case, since momentum space - and particularly the analyticity requirements - rest heavily on the notion of the continuum. Of course, it would not be essential for the continuum to break down at the level at which Chew's ideas apply. However, it is tempting to think that the breakdown of the applicability of the continuum concept to physics ought to occur at the same level as the breakdown of the notions for which it was originally invented, namely space and time. Indeed, if we take this idea seriously, the consequences are rather alarming. For quantum mechanics itself, irrespective of the theory to which it is applied, rests heavily on the notion of the continuum, since at the outset, the array of complex numbers is required for probability amplitudes and for the linear combinations used in the superposition law. Thus, for a theory based on this idea of continuum breakdown to work, a complete reformulation of quantum mechanics in combinatorial

terms would also be required.

Now it might seem that in order to eliminate complex numbers from quantum theory so as to obtain some "discrete" formalism, it would be necessary to alter the theory, not merely reformulate it. However, it appears this may not be so (at least in simple cases), although a somewhat different attitude to the correspondence principle from the one usually expressed seems to be required. I have included, in an appendix, a brief summary of a theory which achieves such a reformulation within the somewhat limited physical context to which it applies. A full account, entitled "Theory of Quantized Directions", will be published in due course. I have not thought it appropriate to include this full account here, although it is very relevant to the above discussion, owing to the fact that the theory is a non-relativistic one. The central idea is to regard total spin as the primary physical concept. Then we can use purely combinatorial rules for obtaining the probability laws governing spin values of systems, and of subsystems into which they split up and recombine. The secondary concept of spin in a direction can be described in terms of total spin, if we regard a "direction" as being something defined by a subsystem of large total spin. In this way, the geometry of directions and the quantum mechanics of angular momentum can be built up simultaneously starting from a purely combinatorial theory. A fully relativistic version of this theory has not yet emerged, however. My own tentative view as to the most promising line of attack would be to combine these ideas with some extensions of the results of sections IV and V and employ some form of "twistor calculus". (Twistor algebra[23] describes the geometry of Minkowski space in a somewhat non-local way, "events" being no longer the primary elements. Twistors can be made also to fit in with general relativity.)

Admittedly, any such combinatorial reformulation of quantum theory would be a major task. But it should not be felt that the successes of "continuous" quantum

theory or, more particularly, the recent successes of techniques of analytic continuations (for scattering amplitudes) in any way imply that such a reformulation necessarily could not work. There would be nothing against analytic techniques being used to obtain important results in a purely combinatorial theory. We need only witness the successes of the Hardy-Littlewood method in number theory to be provided with an example.

To summarize my own view on the form that the next (?) major revolution in the physics of space-time "ought" to take, I should say, then, that some form of combination of (1.2c) and (1.2d) would be expected at some level. It is worth making this point here because, although (needless to say) I do not possess a detailed theory of this type, it serves as a dim and distant goal towards which much of my own research has been directed. This, therefore, supplies some of the motivation for the introduction of certain techniques into relativity theory, which may have been regarded as somewhat "unusual". This applies, particularly to the employment of spinors, rather than tensors, for the description of general relativity, the idea here, being that spinors represent a breakdown of space-time into more elementary parts. Also, the quasi-topological arguments used in section VIII are largely directed at trying to ascertain whether effectively singular states exist in the universe, at which one might expect any breakdown in the manifold structure to be of real significance. I have not attempted any actual modification to Einstein's theory, since I feel this would be definitely premature at this stage. Instead, I have contented myself with certain reformulations and "ways of looking at" the theory in some of its geometrical aspects. The results obtained (one or two of which may be felt to be somewhat surprising) would seem amply to justify the methods used. It is hoped that by reformulating the theory in a sufficiently unusual, yet promising way, one might eventually gain the insight necessary to determine when, whether and how to enlarge the theory, perhaps in accordance with some of the ideas mentioned earlier.

If Einstein had never existed but Newton's gravitational theory had been reformulated as a space-time theory (a very big "but"!), then the discovering of

general relativity would not seem so very great a step, perhaps not even requiring Einstein's genius for its achievement!

II. MATHEMATICS FOR CURVED SPACE-TIME

The picture of space-time as being accurately representable as a differentiable 4-manifold will be effectively assumed in all that follows. However, possibilities for reformulating the theory will be kept in mind, in case different concepts from the ones usually regarded as basic may eventually be able to take over a more fundamental role. I shall not enter into the details of the precise definition of a differentiable manifold here, assuming that the reader is already familiar with this.

Let us, then, consider the way in which the modern differertiable manifold concept of space-time $\mathcal{M}$ (the elements of $\mathcal{M}$ being called events) can be built up[4], indicating this only by general headings:

(2.1a) $\mathcal{M}$ is a topological space (i.e. open subsets of $\mathcal{M}$ defined)

(2.1b) this topological space is Hausdorff and connected

(2.1c) $\mathcal{M}$ has a differentiable structure of dimension 4 (defined in terms of a set of allowable maps from open subsets of $\mathcal{M}$ onto open subsets of the set, R^4, of quadruples of real numbers - or, put another way, in terms of allowable overlapping coordinate neighborhoods)

(2.1d) parameterized curves, smooth or otherwise, defined on $\mathcal{M}$ (as mapping differentiable or merely continuous, of the open interval R^1 into $\mathcal{M}$)

(2.1e) differentiable (scalar) functions defined on $\mathcal{M}$

(2.1f) contravariant vector fields defined on $\mathcal{M}$ (as derivations on the algebra of scalar functions on $\mathcal{M}$)

(2.1g) tangent space defined at each event P (as the affine space of contravariant vectors at P)

(2.1h) covariant vector fields (or 1-forms) defined (as dual of contravariant vector fields)

(2.1i) tensor fields defined (for example, in terms of tensor products of contravariant and covariant vector fields)

(2.1j) a particular non-degenerate, symmetric, covariant tensor field g of valence 2 and of signature -2 is chosen --referred to as the metric tensor-- which defines the "geometry of $\mathcal{M}$

(2.1k) g defines a null cone in the tangent space at each event P

(2.1ℓ) notions of timelike smooth curve and of non-spacelike curve, etc. defined, leading to concept of causality connections between events

(2.1m) measure of time assigned to any timelike smooth curve, as the "distance" assigned by the metric g (direct physical interpretation via the notion of idealized clock); similarly (but somehwat less physically) distance measure assigned to any spacelike smooth curve

(2.1n) notion of covariant differentiation on $\mathcal{M}$, uniquely defined by the requirements that g be covariantly constant and that the torsion should vanish

(2.1o) curvature tensor defined.

At this point we can, if desired, introduce other tensor fields to describe the various (non-gravitational) fields present and construct their stress-energy-momentum tensor T. Finally, we can set up Einstein's field equations by equating T to an expression in the contracted curvature tensor. Modifications of Einstein's theory, such as the unified field theories[9] or the Sciama-Kibble theory[10] require minor modifications at (2.1j) or at (2.1n).

A feature of the way that the theory has been set up is that concepts which are, from the point of view of the mathematics, very elementary and basic may be rather obscure and even irrelevant from the point of view of physics. A prime example of this occurs even at (2.1a). For the essential topological concept of an open set has almost no meaning at all in strictly physical terms. On the other hand, it is difficult to see how one could dispense with this concept without altering the mathematical theory. I shall, however indicate one possible way out which, although the idea will not be developed in full detail here, serves to illustrate a kind of way that one

might hope to reformulate the theory in somewhat more physical terms, yet without changing its mathematical content.

This idea would be to regard the concept of causality, or the related concept of chronology, as more fundamental than the topology of $\mathcal{M}$. In the more usual type of development indicated earlier, these concepts would emerge only at stage (2.1ℓ). I had better be more precise as to the meanings I am assigning to these terms within the usual framework. In fact, these concepts will be needed also in section VIII. If A and B are two events, let us write[29] $A \ll B$ for "A chronologically precedes B", that is to say (loosely) "an observer can travel from A to B; $A \neq B$". Also, write $A \prec B$ for "A causally precedes B", that is, (loosely) "a signal can travel from A to B". The precise meaning of these symbols is:

(2.2a) $A \ll B$ means "there exists a future-oriented timelike smooth parameterized curve containing A and B, for which the parameter value at A is less than that at B";

(2.2b) $A \prec B$ means "there exists a future-oriented non-spacelike smooth parameterized curve containing A and B for which the parameter value at A is not greater than that at B".

A smooth parameterized curve is timelike if its tangent vectors are all timelike and it is non-spacelike if its tangent vectors are all either timelike or null; it is future-oriented if its tangent vectors all lie on or within the "future" null cone at each event on the curve. This notion presupposes that:

(2.3) the null half-cones on $\mathcal{M}$ can be divided into two disconnected classes, which we may label future and past.

That is to say, we suppose that along no closed curve do the null half-cones become interchanged.[30] From the point of view of the physics, this is a very "reasonable" requirement.

Another property that it will also be convenient to assume and which again is very "reasonable" on physical grounds is

(2.4) $\mathcal{M}$ contains no closed non-spacelike smooth curves.

This is equivalent to

(2.5) $A \prec B$ and $B \prec A$ imply $A = B$.

A simple consequence is

(2.6) $A \ll A$ for no event A,

which is equivalent to (2.4) but with "non-spacelike" replaced by "timelike". We can think of (2.6) as expressing the fact that no observer may be allowed to travel into his own past, while (2.5) expresses thefact that a signal should not be able to "travel into its own past".

We define the causal or chronological future and the causal or chronological past of an event A respectively by

(2.7a) $J_+(A) = \{X: A \prec X\}$

(2.7b) $I_+(A) = \{X: A \ll X\}$

(2.7c) $J_-(A) = \{X: X \prec A\}$

(2.7d) $I_-(A) = \{X: X \ll A\}$.

Now, it is easy to see that $I_+(A)$ and $I_-(A)$ are both open sets in the ordinary manifold topology. (If $A \ll X$ and Y lies to the past of X on a timelike curve containing A and X, then X clearly lies in the interior of $I_+(Y)$, whereas $I_+(Y) \subset I_+(A)$.) On the other hand, $J_+(A)$ and $J_-(A)$ need not be closed sets. (To see this, consider Minkowski space-time with a single point removed. A less artificial example is afforded[30a] by a gravitational plane wave.)

We define[29] a topology $\mathcal{T}^*$ on M by taking the sets $I_+(A) \cap I_-(B)$, as A and B vary over M, to be base for the open sets of $\mathcal{T}^*$. (The idea of giving the topology of Minkowski space-time in this way appears to be due to Alexandrov.) For some space-time manifolds, however, $\mathcal{T}^*$ will not be the same as the ordinary manifold topology. But a criterion for the equivalence of the two topologies is the following:

(2.8) There is no event A in $\mathcal{M}$ with the property that every sufficiently small neighborhood of A (in the sense of the manifold topology) is entered twice by some timelike curve (that is, with the curve meeting the neighborhood in a disconnected set).

A statement which is, in fact, equivalent to (2.8), but given purely in terms of causality and chronology is[29]

(2.9) if $A \prec B$ and every event in $I_-(B)$ causally precedes every event in $I_+(A)$, then $A = B$.

This is equivalent, again, to[29]

(2.10) $\mathcal{J}^*$ is Hausdorff.

The proof of equivalence of these statements is not difficult, but will not be given here.

The condition (2.8) may be thought of as implying the non-existence of "almost closed timelike curves". It implies, in particular, that if $I_+(A) = I_+(B)$ or if $I_-(A) = I_-(B)$, then $A = B$. The following example, due to S. W. Hawking, shows that (2.8) (and (2.9), (2.10)) can be violated even though the sets $I_+(A)$ and $I_-(A)$ are all distinct ((2.3) and (2.4) being satisfied): consider the subset of 2-dimensional Minkowski space-time(metric: $ds^2 = (dx^0)^2-(dx^1)^2$ obtained by deleting the portions $(x^0 \leqslant 1,\ x^1 > -1), (x^0 \geqslant -1,\ x^1 < 1)$ and the two points (1,1), (-1,-1); finally identify the two points $(1, x^1)$ and $(-1, -x^1)$ whenever $x^1 > 1$. (The points (1,1) and (-1,-1) at which the manifold is "incomplete" can be sent to infinity by means of a conformal transformation if desired. This will not affect the causal structure.) This example may also be made 4-dimensional by adding two more inessential Minkowski coordinates x^2, x^3 (metric: $ds^2 = (dx^0)^2-(dx^1)^2-(dx^2)^2-(dx^3)^2$. Examples, for which, say, the sets $I_+(A)$ are <u>not</u> all distinct, can also be constructed.[29] If $I_+(A) = I_+(B)$ and $A \neq B$, this implies that timelike curves into the future from A can return arbitrarily near to A.

It seems reasonable to rule out "almost closed timelike curves" on physical grounds, since any space-time which violates (2.8) can, by means of an arbitrarily small perturbation in a compact region, be made also to violate (2.4). If we accept (2.8), then we have a way of defining the manifold topology solely in terms of chronology. Of course, to build the theory up completely, starting from the

chronology concept, it would be necessary to give sufficient axioms for the $\ll$ relation. There is no reason to believe that such an axiom system would be in any way simple, but it does seem that it would be possible to build up the theory in this way.

It may be felt that, on the other hand, the causality concept would be a more natural relation to start from than chronology since fields can transmit signals with speeds up to and including that of light, whereas it is particles or macroscopic observers which are restricted to speeds less than that of light. In fact, we can define chronology in terms of causality in the following way. Let us use the notation

(2.11) $[A, B]$ for $\{X: A \prec X \prec B\}$

and define the relation $\nearrow$ of horismos by

(2.12) $A \nearrow B$ if and only if (i) $A \prec B$

and (ii) $[U, V]$ is linearly ordered by $\prec$ whenever $[U, V]$ is a proper subset of $[A,B]$

Then it turnes out[29] that $\ll$ is just the set-theoretic difference between $\prec$ and $\nearrow$, that is,

(2.13) $A \ll B$ if and only if $A \prec B$ but not $A \rightarrow B$..

Again proofs (which are not difficult) will not be given here, but we have, in principle, then, a way of defining the topology from causality (via chronology).

We may think of the horismos as defining the causal null cone of an event:

(2.14a) $C_+(A) = \{X: \quad A \nearrow X\}$

(2.14b) $C_-(A) = \{X: \quad X \nearrow A\}$.

These sets are generated by null geodesic segments through A. It will turn out, in fact, that

(2.15) $A \nearrow B$ if and only if there is a future-oriented null geodesic from A to B but no timelike curve from A to B.

If we want to regard $\nearrow$ as the fundamental causal relation and to define $\prec$ and $\ll$ in terms of it, we can do this via (2.13), as follows (assuming (2.8) holds):

(2.16) $A \prec B$ <u>if and only if</u> there exists a finite sequence of events $(X_i)_{1 \leq i \leq n}$ satisfying $A = X_1 \nearrow X_2 \nearrow \ldots \nearrow X_n = B$

There could be some physical significance in this sort of thing. We may conceive that all interactions may ultimately be broken down into effects which propagate with the speed of light, any effects propagating more slowly than that of light being built up of "null zig-zags" in the manner of (2.16). This is somewhat reminiscent of the "zitterbewegung" of the electron, whereby a net timelike motion of the particle is built up, although its velocity is always instantaneously that of light. This point of view, in fact, fits in well with the ideas of Section IV, but it will not be pursued further here.

It is not my purpose, here, to discuss in detail such possible reformulations of general relativity theory. I have been able, in the above, only to touch on the first stages (2.1a), (2.1b) in the construction of a manifold. To continue to develop the ideas, so as to encompass all stages up to (2.1ℓ), would surely be a major task. I do not even want to suggest that I think this would necessarily be a fruitful thing to do (although I think it would at least be interesting). I am somewhat undecided, for example, as to the virtue of incorporating <u>global</u> assumptions about the nature of space-time (cf. (2.4), (2.8)) at the outset. But I feel that there is at least one feature of significance in the above type of approach, namely the elevation of the importance of the null cone structure of space-time over and above that of its manifold (i.e. continuity - differentiability) structure. As an illustration of why I feel this way, consider the picture conjured up by a position space attitude to quantum field theory. As we localize the position of a particle it jumps essentially along the null cone. Other particles are produced, which leap backwards and forwards essentially along null directions, without apparent regard for continuity, heeding only the positions of the null cones themselves and "topology" only in the respect in which this term is applied to the structure of

graphs. The apparently "tame" and continuous appearance of macroscopic phenomena results from a superposition of many chaotic pictures of this kind. But perhaps the position space attitude to events in the small is a misleading one? Nonetheless, however one formulates quantum field theory, the fundamental importance of a local null cone structure as afforded by local Poincaré covariance can hardly be doubted.

Let me put things another way. In the development of the relativistic manifold concept outlined in (2.1), the concept of a metric tensor was introduced in (2.1j) by simply choosing a particular type of tensor field g and asserting that this defines the "geometry" of $\mathcal{M}$. But from the point of view of an abstract differentiable manifold there is nothing special about a symmetric covariant tensor field of valence 2. Indeed, why should we not introduce all the other fields of nature on exactly the same footing? For example, we might introduce a particular skew-symmetric, covariant tensor field F of valence 2 (namely the Maxwell field tensor) and assert that F, also, must be referred to as "geometry".[31] It is true that g directly defines the way in which (idealized!) clocks measure time and hence it also defines the physical measurement of distance. But F, too, strongly influences physical objects, and in particular the motions of particles. What then, is the real difference between g and "other" physical fields, such as F, which requires that g must be introduced as a primary concept - as the essential part of the very structure of space-time itself, whereas F, etc. can be regarded as somewhat incidental objects which we may be inclined to think of, rather, as residing "in" space-time? Part of the answer, surely, is contained in the universal nature of the field g. (The principle of equivalence is, perhaps, one aspect of this universality.) When one starts to write field equations (or Lagrangians) for the other fields, it always emerges that the g field must also enter in an essential way - and it enters into the free-field terms.[32] If any other field were to enter in this way, we would assert that this represents, instead, an interaction. But could it be that the attitude here is wrong? Ought we to redefine the concept of interaction so that what we normally refer to as a "free-field" would instead have to be thought of as a field continually

interacting with g? Undoubtedly not. The free-field concept itself would then lose all meaning.

An intermediate point of view (and this, effectively, is the one normally adopted for the study of weak gravitational fields) is to introduce a flat Minkowskian "background" metric η, to write $g = \eta + h$, then to regard the free-field part as that involving η but not h, and the(gravitational) interaction terms as the ones involving h. But this is a highly arbitrary (i.e. "non-covariant") procedure, since there is, in general, no unique way of defining the η metric. The only way to preserve complete covariance (or, what we may call the "geometrical attitude") while retaining the normal notion of a free field in the absence of gravitation, is to allow the gravitational interaction to find expression through the curvature tensor. This it does via the non-commutativity of covariant derivatives (cf. 3.23, 3.29). The gravitational field is thus described, not directly by the tensor field g, but by the curvature tensor (or, more conveniently, by the Weyl curvature tensor, see 3.27). This attitude also has the advantage that it brings out the most striking of the formal analogies between gravitation and electromagnetism (see, for example, Sections IV, V, VI).

These, then, are some of the reasons why the tensor field g, and the associated system of null cones, must be regarded as being on quite a different footing from other physical fields, as constituting, in fact, the fundamental geometric structure of $\mathcal{M}$. But there is also an additional reason for this - one which, in my view, is quite conclusive. For, so long as we restrict ourselves to fields of integral spin, we have this possibility of regarding these fields as tensorial quantities on a "structureless" differentiable manifold $\mathcal{M}$, which only incidentally (apparently), via field equations or special relations, are related to a certain other field g. But with spinor fields, this possibility is no longer open to us. Without a pre-existing[33] metric on $\mathcal{M}$, we cannot even introduce the concept of a spinor.[34] The basic importance of spinors to physics can hardly be doubted; and spinors are quantities essentially defined in conjunction with the null cones, not with the abstract differential geometry structure of $\mathcal{M}$.

The argument may also be put the other way. We may ask whether there can be an "explanation" of the special role of g; indeed, whether the particular dimensionality and signature of our space-time - which seem so arbitrary from the point of view of the mathematical development of a manifold - can be built up from some more "elementary" structure, and in a way which perhaps removes some of this arbitrariness. To my mind, there can be no doubt but that the existence of spinors must be deeply related to an answer to this question. For, while it is true that spinors of a kind can be defined for <u>any</u> dimensionality and signature,[35] the role that can be played by such spinors differs, somewhat, in all other cases, from the special role enjoyed by the two-component spinors appropriate to our space-time, with its $(+,-,-,-)$ structure. As the most elementary illustration of this fact, the general (2x2) Hermitian determinant is a quadratic form of Lorentzian signature in four real variables :

$$\begin{vmatrix} t+z & x+iy \\ x-iy & t-z \end{vmatrix} \equiv t^2 - x^2 - y^2 - z^2 \tag{2.17}$$

This has no direct analogue for any other dimension or signature.

The idea here, then, is to attempt to exploit to the full this special connection between two-component spinors and the light-cone structure of space-time manifolds. The general philosophy here is strangely opposite to that which is normally fruitful in pure mathematics. In pure mathematics, one tries to exploit formalisms of maximum generality, in the hope that results which are useful in one context may also be employed in others, perhaps far removed from the original applications, thereby obtaining a broader and more unified outlook. But here, the very generality of a formalism may, to some extent, be an indication that one is partly "on the wrong track". For, if a formalism enables one to treat myriads of non-existent types of universe, then (effectively) it contains "arbitrary parameters", only special values of which will correspond to world as it actually is. In the ordinary approach to space-time as a pseudo-Riemannian differentiable manifold, the dimension of the manifold and signature of the metric

are two such arbitrary parameters. The hope is that by concentrating on a formalism which is tailored to the particular structure of our space-time, we may eventually learn more about this structure; and if it finally becomes necessary to change the formalism so as to fit in with a new view of the nature of space-time, then we shall have a suitable vantage point from which to achieve this.

Let us return to the identity (2.17) and extract its most immediate implications. We are interested in linear transformations of t, x, y, z which preserve the right-hand side of (2.17), since these define the Lorentz group - the local symmetry group of space-time. Now a transformation

$$\begin{pmatrix} t'+z' & x'+iy' \\ x'-iy' & t'-z' \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} t+z & x+iy \\ x-iy & t-z \end{pmatrix} \begin{pmatrix} \bar{\alpha} & \bar{\gamma} \\ \bar{\beta} & \bar{\delta} \end{pmatrix}. \tag{2.18}$$

$\alpha, \beta, \gamma, \delta$ being complex numbers, clearly preserves Hermiticity (i.e. the reality of t,x,y,z) and also the determinant (2.17), provided

$$\begin{vmatrix} \alpha & \beta \\ \gamma & \delta \end{vmatrix} \equiv \alpha\delta - \beta\gamma = 1 \tag{2.19}$$

Thus, (2.18), (with (2.19)), is a Lorentz transformation. It is, in fact the most general Lorentz transformation continuous with the identity. This follows by direct construction or from the fact that transformations of the form (2.18), satisfying (2.19), form a six-dimensional system. This latter fact follows from the property that only a discrete set of values of $\alpha, \beta, \gamma, \delta$ can give the identity transformation in (2.18). Indeed, one readily obtains exactly the <u>two</u> sets of values. $\alpha=\delta=1$, $\beta=\gamma=0$ and $\alpha=\delta=-1$, $\beta=\gamma=0$, which satisfy (2.19) and give the identity transformation in (2.18). It follows that to <u>any</u> Lorentz transformation there correspond precisely <u>two</u> $(\alpha,\beta;\gamma,\delta)$ matrices - referred to hereafter as <u>spin-matrices</u>. From this, and the fact that the spin-matrices form a connected manifold, follows the fundamental result[35] that the spin-matrices form a two-valued faithful representation of the proper orthochronous[36] Lorentz group.

Thus, we may represent components of world vectors as Hermitian (2x2) matrices, rather than "linearly":

$$\frac{1}{\sqrt{2}}\begin{pmatrix} t+z & x+iy \\ x-iy & t-z \end{pmatrix} \longleftrightarrow (t, x, y, z) \tag{2.20}$$

(the factor $\sqrt{2}$ is for later convenience). The matrices (2.20) transform according to the direct product representation of the spin-matrix group with its complex conjugate (cf. (2.18)). But what are the spin-vectors themselves? That is, how are we to interpret the quantities which are transformed directly by spin-matrices? In some sense they are "square-roots" of vectors, but what does this mean precisely? As the first clue, we observe that when the invariant (2.17) vanishes, the matrix in (2.20) has rank unity. Thus (if $t > 0$) it is the outer product of a complex column 2-vector with its Hermitian conjugate. This 2-vector is really a spin-vector since we can allow it to be transformed by spin-matrices under Lorentz transformation. There is a phase factor which is not defined by this construction, but apart from this, we see that it is the null (future-pointing) world vectors which define spin vectors. In fact the phase of the spin vector can also be interpreted[37] (see 3.20) as a half-plane element, which both contains and is orthogonal (in the Lorentzian sense) to the direction of the null vector. Then we have an almost complete geometrical picture of a spin vector as a kind of "null flag" in space-time.

The one respect in which a picture of this kind must necessarily fail to be complete arises from the essential two-valuedness of the spin-matrix representation of the Lorentz group. A continuous sequence of Lorentz transformations representing the successive stages in a rotation (about some axis) through 2π must, if represented continuously by spin-matrices, correspond to a sequence which starts with the identity spin matrix but ends with the negative identity spin matrix (or vice-versa). Thus, if a spin-vector is rotated continuously about some axis, through an angle 2π, we end up, not with the original spin-vector, but with its negative. If we wish for a complete (necessarily somewhat non-local) representation of a spin-vector in

space-time, we must have a way of keeping track of the (parity of the) number of times it has been rotated. The net effect of this, from the point of view of the mathematics, is that a spin-vector cannot accurately be represented as a structure in the tangent space at an event on $\mathcal{M}$. Instead, a spin-vector field is a "cross-section of a double covering of the null-flag bundle of $\mathcal{M}$". In fact there is a topological restriction which must be placed on the manifold $\mathcal{M}$ (vanishing of the Stiefel-Whitney class w_2) in order that spin-vector fields be globally definable[38] on $\mathcal{M}$ (i.e. that the null-flag bundle should possess the appropriate double covering). Even if this topological restriction is satisfied, we have trouble if $\mathcal{M}$ is not simply-connected, since then its spin-structure is not uniquely defined by its geometry.

Because spin-vectors have this strange non-local behaviour, it might be argued that they are not really "geometrical" objects and perhaps, therefore, not suitable to be the basic building blocks of a purely geometrical theory, ie. of the structure of space-time. It should be emphasized, however, that this behaviour under rotation is a real part of physics, since it is essential for the consistency of the description of half-odd spin particles. Furthermore it is not, as has sometimes been suggested merely an unobservable aspect of the quantum description of phenomena - the phase of the wave function being "unobservable". To illustrate this fact I shall consider an object[39] which I shall refer to as an "Aharanov box". This consists of two halves which can be fitted together in a particular relative orientation. When together, a current is observed to flow from one half to the other. The two halves are then separated and one half is rotated through an angle 2π relative to the other half. They are placed together again in precisely the same orientation as before and a current is again observed to flow from one half to the other, but now, miraculously, the direction of flow is exactly the reverse of that in the previous case! Only after a further rotation of one half of the box through an angle 2π (about any axis) is the original direction of the current restored. Even though an Aharanov

box has never actually been built and its physical construction would be exceedingly difficult to achieve practice, the essential point is that it is theoretically possible[39] to construct such a box, there being no technological obstacle of principle to be overcome. (Indeed in the Josephson bridge, some of the related practical problems have already been overcome). As is to be expected, the principle of an Aharanov box depends on quantum effects, but these quantum effects add together to produce an observable macroscopic current. When the boxes are in position, the geometry (in the ordinary sense) of the situation does not determine the physical behaviour. But if we extend the concept of geometry to include spinorial quantities, this problem is thereby removed!

I am taking the view, therefore, that whereas the metric structure of space-time may, in some sense, be more fundamental than its continuity-differentiability structure, then also its spinor structure may be even more fundamental than its metric structure. There is, however, a further point which could lead to a slight modification of this view and which certainly should be kept in mind. This is that the really physically important role played by g is simply to define the null cones, and hence give rise to the causality structure of $\mathcal{M}$ (cf. (2.1k), (2.1ℓ)). This requires g to be known only up to proportionality. Thus, it may be that it is actually the conformal structure of $\mathcal{M}$ that is fundamental while the scaling that finally defines the g-field has more the nature of a "physical field" like F (or the "curvature tensor field"). If this attitude is to be entertained seriously from the point of view of physics, then the possibility should also be considered of different coexisting scalings for the g. This would lead to a theory of the Jordan-Thiry-Dicke[41] general type, with a variable gravitational "constant". I shall not consider such theories here, however.

On the other hand, there are certain mathematical advantages in studying space-time from the point of view of its conformal structure (see Section VII). These arise essentially from the fact that the appropriate field equations for fields of zero rest-mass are conformally invariant. Furthermore, for suitable space-times (for example, those which are asymptotically flat), conformal transformations may be used to "make infinity finite"; that is, a boundary hypersurface can be added "at infinity" so that the resulting manifold-with-boundary has a well-behaved conformal structure. This gives a useful

and covariant treatment of the asymptotic properties of zero rest-mass fields (gravitation included).

III. TWO-COMPONENT SPINOR FORMALISM

The expression (2.20) defines, in a special reference frame, the representation of a world-vector as a Hermitian two-index spinor. For a more general reference frame in which the components of the metric tensor need not have the explicit form diag $\{1,-1,-1,-1\}$, we can write (2.20) as

$$X^{AB'} \longleftrightarrow x^{\mu} \tag{3.1}$$

where[40] the elements of the (2x2) Hermitian matrix $(X^{AB'})$ are related linearly to the components (x^{μ}) of a world vector:

$$X^{AB'} = \sigma_{\mu}{}^{AB'} x^{\mu} \tag{3.2}$$

and inversely

$$x^{\mu} = \sigma^{\mu}{}_{AB'} X^{AB'} \tag{3.3}$$

Here, $(\sigma_{\mu}{}^{AB'})$ and $(\sigma^{\mu}{}_{AB'})$ are Hermitian (2x2) matrices for each μ. (The summation convention is being used consistently here. Greek letters are used for tensor indices, running over 0,1,2,3. Spinor indices are represented by upper or lower capital Roman letters, which are unprimed or primed according as they correspond to transformations by spin matrices or by complex conjugate spin matrices. They run over two values 0,1 or 0',1' respectively.) Now the fundamental relation (2.17) becomes

$$2 \left| X^{AB'} \right| = g_{\mu\nu} x^{\mu} x^{\nu} \tag{3.4}$$

where $(g_{\mu\nu})$ are the components of metric tensor g. Thus

$$\begin{aligned} g_{\mu\nu} x^{\mu} x^{\nu} &= \epsilon_{AC} \epsilon_{B'D'} X^{AB'} X^{CD'} \\ &= \epsilon_{AC} \epsilon_{B'D'} \sigma_{\mu}{}^{AB'} \sigma_{\nu}{}^{CD'} x^{\mu} x^{\nu} \end{aligned}$$

by (3.2). Here ϵ_{AC}, $\epsilon_{B'D'}$ are, together with ϵ^{AC}, $\epsilon^{B'D'}$, the spinor Levi-Civita symbols (skew-symmetric expressions with $\epsilon_{01}=\epsilon_{0'1'}=\epsilon^{01}=\epsilon^{0'1'}=1$). Hence

$$g_{\mu\nu} = \epsilon_{AC}\,\epsilon_{B'D'}\,\sigma_{\mu}{}^{AB'}\sigma_{\nu}{}^{CD'} \tag{3.5}$$

(by the arbitrariness of x^{μ} and the symmetry of both sides of (3.5) in μ and ν). Inversely, we have

$$\epsilon_{AC}\,\epsilon_{B'D'} = g_{\mu\nu}\,\sigma^{\mu}{}_{AB'}\,\sigma^{\nu}{}_{CD'} \tag{3.6}$$

The general spinor-tensor correspondence is exemplified by

$$\Upsilon^{AB'\,CD'}{}_{EF'} = \sigma_{\lambda}{}^{AB'}\,\sigma_{\mu}{}^{CD'}\,\Upsilon^{\lambda\mu}{}_{\nu}\,\sigma^{\nu}{}_{eD'}\,\epsilon^{e}{}_{E}\,\epsilon^{D'}{}_{F'} \tag{3.7}$$

where, inversely,

$$\Upsilon^{\lambda\mu}{}_{\nu} = \sigma^{\lambda}{}_{AB'}\,\sigma^{\mu}{}_{CD'}\,\Upsilon^{AB'\,CD'}{}_{EF'}\,\sigma_{\nu}{}^{EF'} \tag{3.8}$$

Writing this $\Upsilon^{AB'\,CD'}{}_{EF'} \longleftrightarrow \Upsilon^{\lambda\mu}{}_{\nu}$ (cf. (3.1),(3.2),(3.3)), we see that the fundamental relation (2.17) finds expression as

$$\epsilon_{AC}\,\epsilon_{B'D'} \longleftrightarrow g_{\mu\nu} \tag{3.9}$$

(see (3.5), (3.6)). Thus the spinor translation of the metric tensor has the important property that it "splits" into the two Levi-Civita "metrics", for spin space and for the complex conjugate spin space.

In fact (3.9) contains all the essential information that is required for the formal spinor treatment of local space-time. It is not necessary ever to consider the σ's explicitly. Also, I shall not enter into the subtleties here, concerning the mathematical definition of spinor fields on the manifold $\mathcal{M}$, which properly takes into account their two-valued nature. It will be sufficient to consider the formal algebra of the spinors and their relation to the metric as given by (3.9). In fact, I shall tend to regard quantities such as ϵ_{AB}, $g_{\mu\nu}$, $\Upsilon^{\lambda\mu}{}_{\nu}$, ξ^{A}, not so much as sets

of components, but as algebraic-geometric objects in their own right, subject to certain formal rules of manipulation. The sole purpose of the σ's, then, is to enable us to treat any tensor index formally as though it were a pair of spinor indices, one unprimed and one primed.

The ϵ's will be used for raising and lowering spinor indices:

$$\xi_{\dots\ \dots}^{\dots A \dots} = \epsilon^{AB}\xi_{\dots\ B\dots}^{\dots\ \dots}\ , \quad \xi_{\dots B\dots}^{\dots\ \dots} = \xi_{\dots\ \dots}^{\dots A\dots}\epsilon_{AB} \tag{3.10a}$$

$$\eta_{\dots\ \dots}^{\dots A'\dots} = \epsilon^{A'B'}\eta_{\dots B'\dots}^{\dots\ \dots}\ , \quad \eta_{\dots B'\dots}^{\dots\ \dots} = \eta_{\dots\ \dots}^{\dots A'\dots}\epsilon_{A'B'} \tag{3.10b}$$

(note ordering of indices). It follows from (3.5) and (3.6) (using $\sigma^{\mu}{}_{AB'}\sigma_{\nu}{}^{AB'} = \delta^{\mu}_{\nu}$ or $\sigma_{\mu}{}^{AB'}\sigma^{\mu}{}_{CD'} = \delta^{A}_{C}\delta^{B'}_{D'}$) that (3.10) is consistent with the notation for the σ's and with

$$Z_{\dots\ \dots\mu\dots}^{\dots\ \dots} = Z_{\dots\ \dots}^{\dots\nu\dots}g_{\nu\mu}\ , \quad Z_{\dots\ \dots}^{\dots\mu\dots} = Z_{\dots\ \dots\nu\dots}^{\dots\ \dots}g^{\mu\nu} \tag{3.11}$$

($g^{\mu\nu}$ being inverse to $g_{\mu\nu}$), i.e. $\sigma_{\mu}{}^{AB'} = \epsilon^{AC}\epsilon^{B'D'}\sigma^{\nu}{}_{CD'}\,g_{\nu\mu}$.

Unprimed spinor indices may be contracted only with unprimed indices and primed only with primed. Note also that

$$\zeta_{\dots\ \dots\ \dots}^{\dots A\dots\ \dots} = -\zeta_{\dots\ \dots A\dots}^{\dots\ \dots A\dots}\ , \quad \theta_{\dots\ \dots\ \dots}^{\dots A'\dots\ \dots} = -\theta_{\dots\ \dots A'\dots}^{\dots\ \dots A'\dots} \tag{3.12}$$

$$\epsilon_{A}{}^{B} = \delta_{A}^{B} = -\epsilon^{B}{}_{A}\ , \quad \epsilon_{A'}{}^{B'} = \delta_{A'}^{B'} = -\epsilon^{B'}{}_{A'} \tag{3.13}$$

Furthermore, $\epsilon_{AB}\epsilon^{CD} = \delta_{A}^{C}\delta_{B}^{D} - \delta_{A}^{D}\delta_{B}^{C}$, whence

$$\zeta_{\dots AB\dots} - \zeta_{\dots BA\dots} = \zeta_{\dots\ C\dots}^{\ \ C}\epsilon_{AB}\ . \tag{3.14}$$

The ϵ's satisfy

$$\epsilon_{AB}\epsilon_{CD} + \epsilon_{AC}\epsilon_{DB} + \epsilon_{AD}\epsilon_{BC} = 0 \tag{3.15}$$

which is equivalent to (3.14). Also the alternating tensor

(3.16) $$e_{\mu\nu\rho\sigma} = (-g)^{\frac{1}{2}} \epsilon_{\mu\nu\rho\sigma}$$

has a spinor translation involving only the spinor ϵ's. Using the form $e^{\mu\nu}{}_{\rho\sigma}$ which is useful for "dualizing", we can express this as[41]

(3.17) $$e^{AB'CD'}{}_{EF'GH'} = i\delta^A_E\delta^C_G\delta^{B'}_{H'}\delta^{D'}_{F'} - i\delta^A_G\delta^C_E\delta^{B'}_{F'}\delta^{D'}_{H'}$$

We may verify (3.17) up to sign by noting that it gives (via (3.15)) a completely skew-symmetrical $e_{\mu\nu\rho\sigma}$ for which $e_{\mu\nu\rho\sigma}e^{\mu\nu\rho\sigma} = 24$. (In fact, the sign of (3.17) depends on the particular choice of σ's which is made.) Any tensor relation can thus be interpreted as an equivalent spinor relation, tensor contractions translating into spinor contractions.

Spinors are naturally <u>complex</u> objects, the notion of reality of a spinor component having no invariant meaning (except in the case of scalars). Reality of a tensor is interpreted, instead, as a <u>Hermitian</u> property of the corresponding spinor, the complex conjugate of an unprimed index being a primed index and vice versa. For example, the complex conjugate of $\Upsilon^{AB'CD'}{}_{EF'}$ is written $\overline{\Upsilon}^{A'BC'D}{}_{E'F}$. The condition that $\Upsilon^{\mu\nu}{}_{\rho}$ be real is then the Hermiticity relation $\Upsilon^{AB'CD'}{}_{EF'} = \overline{\Upsilon}^{B'AD'C}{}_{F'E}$. If $g_{\mu\nu} = \mathrm{diag}\{1,-1,-1,-1\}$, a convenient representation for $2^{1/2}\sigma^{\mu}{}_{AB'}$, satisfying (3.6), is a unit matrix together with the Pauli matrices. The relation between a vector a^{μ} and its spinor equivalent $a^{AB'}$ is then just (2.20), i.e.

(3.18a) $$\begin{pmatrix} a^{00'} & a^{01'} \\ a^{10'} & a^{11'} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} a^0 + a^1 & a^2 + ia^3 \\ a^2 - ia^3 & a^0 - a^1 \end{pmatrix};$$

(3.18b) $$(a^0, a^1, a^2, a^3) = \tfrac{1}{\sqrt{2}}(a^{00'} + a^{11'}, a^{00'} - a^{11'}, a^{10'} + a^{01'}, ia^{10'} - ia^{01'}).$$

Now let us consider the <u>geometrical interpretation</u>[37] of a one-index spinor ξ_A. The vector u_{μ} whose spinor equivalent is given by

CHAPTER 28

(3.19) $$u_{AB'} = \xi_A \bar{\xi}_{B'}$$

is real, null and (with the appropriate choice for the sign of $\sigma^{\mu}{}_{AB'}$) future-pointing. (That u_μ is null follows from $\xi_A \xi^A = 0$. Note, generally, that $\xi_A \zeta^A = 0$ if and only if ζ_A is a scalar multiple of ξ_A.) Conversely, any such null vector can be expressed in the form (3.19). However ξ_A contains more information than u_μ since multiplication of ξ_A by a phase factor does not affect (3.19). But the phase of ξ_A also has a geometrical interpretation. The real bivector $p_{\mu\nu}$ with spinor equivalent

(3.20) $$p_{AB'CD'} = \xi_A \xi_C \epsilon_{B'D'} + \epsilon_{AC} \xi_{B'} \xi_{D'},$$

constructed from the square of ξ_A, has the form

(3.21) $$p_{\mu\nu} = u_\mu w_\nu - u_\nu w_\mu$$

where

(3.22) $$w_{AB'} = \xi_A \bar{\eta}_{B'} + \eta_A \bar{\xi}_{B'}, \quad \xi_A \eta^A = 1,$$

as is readily verified. The vector w_μ is real and spacelike (with $w_\mu w^\mu = -2$), orthogonal to u_μ, and is determined by ξ_A up to the addition of a multiple of u_μ. Thus ξ_A determines, not only the null vector u_μ, but also a certain null half-plane element containing the null direction, namely, the half-plane of directions of $w_\mu + \lambda u_\mu$. A spinor may therefore be conveniently pictured as a kind of "flag" (see fig. 1). The flagpole represents the null vector u_μ and the flag itself the half-plane element determined by u_μ and w_μ. This flag and flagpole gives a complete geometrical realization of ξ_A apart from a sign ((3.19) and (3.20) are invariant under $\xi_A \to -\xi_A$). This sign ambiguity is, of course, essential for such a direct ("local") geometrical interpretation. A continuous spacial rotation through an angle 2π changes ξ_A continuously into $-\xi_A$.

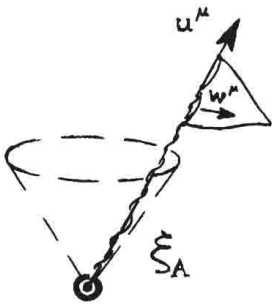


Fig. 1. A spinor ξ_A is represented geometrically, up to a sign, as a "flag". The flagpole is a null vector u_μ and the flag itself is a half-plane element spanned by u_μ and a spacelike vector w_μ orthogonal to it. To distinguish the representation from that of a primed spinor $\zeta_{B'}$ a screw sense can be associated with the flagpole.

Note that the null direction of the flagpole is unchanged if ξ_A is multiplied by any non-zero complex scalar λ. The "extent" of the flagpole is, however, multiplied by a factor $|\lambda|^2$ and the flag rotates through an angle 2 arg λ. Two complete rotations of the flag thus are necessary to correspond to a phase change through an angle 2π for ξ_A. The representation of a spinor $\zeta_{B'}$ with a single primed index is essentially the same as that for ξ_A except that the flag rotates in the opposite direction when $\zeta_{B'}$ is multiplied by λ. One can therefore associate a screw sense with a spinor ξ_A and associate the opposite screw sense with $\zeta_{B'}$. This is related to the spin properties of zero rest-mass particles.

It is conceptually very useful that there is a clear-cut geometrical representation of spinors. For it removes most of the abstractness that is often associated with "spin space". In fact, it is only the sign ambiguity which prevents spinors from being just as good geometrical objects (in the usual sense) as tensors. The notion of covariant derivative (denoted by ∇_μ or by $\nabla_{AB'} \equiv \sigma^\mu{}_{AB'} \nabla_\mu$) now has a natural unambiguous definition in terms of parallel displacement. The covariant derivatives of $g_{\mu\nu}$, $\sigma^\mu{}_{AB'}$, ϵ_{AC}, $\epsilon_{B'D'}$ are then all required to be zero. This

fixes the spinor affine connections[64], $\Gamma^{A}_{B\mu}, \overline{\Gamma}^{A'}_{B'\nu}$ in terms of coordinate derivatives of $\sigma^{\mu}_{AB'}$, but details of such coordinate dependent quantities will be left until Section VI, the formal properties of $\nabla_{AB'}$ being all that will be required here.

In pseudo-Riemannian space-time the commutator of two covariant derivatives gives rise to the curvature tensor $R_{\rho\sigma\mu\nu}$:

$$(3.23) \qquad (\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu) k_\sigma = k^\rho R_{\rho\sigma\mu\nu}$$

The spinor translation of $R_{\rho\sigma\mu\nu}$ can be written[41] as follows:

$$(3.24) \qquad R_{AE'BF'CG'DH'} = \Psi_{ABCD} \epsilon_{E'F'} \epsilon_{G'H'} + \epsilon_{AB} \epsilon_{CD} \overline{\Psi}_{E'F'G'H'}$$

$$+ 2\Lambda \{ \epsilon_{AC} \epsilon_{BD} \epsilon_{E'F'} \epsilon_{G'H'} + \epsilon_{AB} \epsilon_{CD} \epsilon_{E'H'} \epsilon_{F'G'} \}$$

$$+ \epsilon_{AB} \Phi_{CDE'F'} \epsilon_{G'H'} + \epsilon_{CD} \Phi_{ABG'H'} \epsilon_{E'F'} ,$$

where

$$(3.25) \qquad \Phi_{ABC'D'} = -\tfrac{1}{2} R_{AC'BD'} + \tfrac{1}{8} R \epsilon_{AB} \epsilon_{C'D'}$$

represents the trace-free part of $R_{\mu\nu}$ $(= R^{\alpha}{}_{\mu\alpha\nu})$,

$$(3.26) \qquad \Lambda = \frac{1}{24} R$$

($R = R^{\alpha}{}_{\alpha}$) describes the scalar curvature, and where Ψ_{ABCD} represents Weyl's conformal tensor $C_{\mu\nu}{}^{\rho\sigma}$ ($= R_{\mu\nu}{}^{\rho\sigma} - 2R_{[\mu}{}^{[\rho}\delta_{\nu]}^{\sigma]} + \frac{1}{3} R \delta_{[\mu}^{\rho} \delta_{\nu]}^{\sigma}$) according to

$$(3.27) \qquad C_{AE'BF'CG'DH'} = \Psi_{ABCD} \epsilon_{E'F'} \epsilon_{G'H'} + \epsilon_{AB} \epsilon_{CD} \Psi_{E'F'G'H'}$$

(Square brackets are used to denote "totally skew-symmetric part" and round brackets, to denote "totally symmetric part.) The tree quantities Ψ_{ABCD}, $\Phi_{ABC'D'}$, Λ determine the parts of $R_{\mu\nu\rho\sigma}$ irreducible under local Lorentz transformations, having the symmetries

(3.28) $$\Psi_{(ABCD)} = \Psi_{ABCD}\,,\quad \Phi_{(AB)(C'D')} = \Phi_{ABC'D'} = \overline{\Phi}_{C'D'AB}\,,\quad \Lambda = \overline{\Lambda}\,.$$

The Ricci identities (3.23) become, in spinor form,[41]

(3.29) $$\left(\nabla_A^{P'}\nabla_B^{Q'} - \nabla_B^{Q'}\nabla_A^{P'}\right)\xi_C = \xi^X\left(\Psi_{XCAB}\epsilon^{P'Q'} + 2\Lambda\,\epsilon_{X(A}\epsilon_{B)C}\epsilon^{P'Q'} + \Phi_{XC}^{P'Q'}\epsilon_{AB}\right).$$

(We can obtain (3.29) from (3.23) applied to (3.20).) Equivalently,

(3.30a) $$\nabla_{(A}^{Y'}\nabla_{B)Y'}\xi_C = \xi_X\,\Psi^X{}_{ABC} + 2\Lambda\;\xi_{(A}\epsilon_{B)\,C}$$

(3.30b) $$\nabla^{Y(P'}\nabla_Y^{Q')}\xi_C = \xi_X\,\Phi_C{}^{XP'Q'}$$

The complex conjugates of (3.29), (3.30) must also be considered. Spinors with many indices are treated by repeated applications of these rules. That is, they behave formally as a product of one-index spinors treated according to the Leibnitz rule. The Bianchi identity $\nabla_{[\tau}R_{\rho\sigma]\mu\nu} = 0$ becomes[41]

(3.31) $$\nabla_{P'}^{A}\Psi_{ABCD} = \nabla_B^{Q'}\Phi_{CDQ'P'} - 2\epsilon_{B(C}\nabla_{D)P'}\Lambda$$

or equivalently,

(3.32) $$\nabla_{P'}^{A}\Psi_{ABCD} = \nabla_{(B}^{Q'}\Phi_{CD)Q'P'}\,,\quad \nabla_{Q'}^{B}\Phi_{AB}^{P'Q'} = 3\nabla_A^{P'}\Lambda$$

In flat space-time, with the usual definition of ∇_μ, the commutator (3.23) of course, vanishes (as do (3.29), (3.30)). However, it is sometimes convenient to consider other possible definitions for ∇_μ. In particular, if α_μ is the electromagnetic potential and ∂_μ is the usual flat space-time gradient operator, we can define

(3.33) $$\nabla_\mu\theta_{\cdots}^{\cdots} \equiv \left(\partial_\mu - ie\alpha_\mu\right)\theta_{\cdots}^{\cdots}$$

where the integer e denotes the <u>charge</u>[42] of the field $\theta_{\cdots}^{\cdots}$. Any spinor must then have

a charge value assigned to it. Spinors may only be added together if they have the same charge value; the charge of the product of two spinors is the sum of their charges; contraction or differentation does not affect the charge; complex conjugation changes the sign of the charge; the ϵ spinors have zero charge. The phase of a spinor then only has geometrical maaning if the charge is zero. (Note that (3.20) is not permissible if ξ_A has non-zero charge.)

This ensures that α_μ, as such, has no geometrical meaning assigned to it. However, from (3.33),

$$(3.34)\qquad (\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu)\,\theta^{\cdots}_{\cdots} = ie F_{\mu\nu}\,\theta^{\cdots}_{\cdots}$$

where $F_{\mu\nu} = \partial_\nu \alpha_\mu - \partial_\mu \alpha_\nu$ is the electromagnetic field tensor. We thus have $F_{\mu\nu}$ defined operationally ("geometrically") by (3.34) (∇_μ is here regarded as the "geometrical" operation, not ∂_μ). In spinor terms we can write

$$(3.35)\qquad F_{AC'BD'} = \phi_{AB}\epsilon_{C'D'} + \epsilon_{AB}\bar{\phi}_{C'D'}$$

with $\phi_{AB} = \phi_{BA}$. Then[43] ϕ_{AB} uniquely represents $F_{\mu\nu}$ with

$$(3.36)\qquad \phi_{AB} = \partial^{C'}_{(A}\alpha_{B)C'}$$

and we can express (3.34) as

$$(3.37a)\qquad \nabla^{Y'}_{(A}\nabla_{B)Y'}\,\theta^{\cdots}_{\cdots} = -ie\phi_{AB}\,\theta^{\cdots}_{\cdots},$$

$$(3.38b)\qquad \nabla^{Y(P'}\nabla^{Q')}_{Y}\,\theta^{\cdots}_{\cdots} = -ie\bar{\phi}^{P'Q'}\theta^{\cdots}_{\cdots}.$$

To treat electromagnetism in curved space-time, we need only combine (3.37) with (3.31), (3.32). For example, if $\theta^{P'Q'}_{A}$ is symmetric in P', Q' and has charge 3 we have

$$(3.38)\qquad \nabla^{R'}_{(C}\nabla_{D)R'}\,\theta^{P'Q'}_{A} = \theta^{P'Q'}_{B}\Psi_{ACD}{}^{B} + 2\Lambda\theta^{P'Q'}_{(C}\epsilon_{D)A} - 2\,\theta^{S'(P'}_{A}\Phi_{CDS'}{}^{Q')} - 3i\phi_{CD}\theta^{P'Q'}_{B}$$

It is also possible to incorporate torsion into the formalism.[10] One further term then appears on the right hand side of (3.38), which involves the derivative of $\theta^{\cdots}_{\cdots}$ and a torsion spinor.

IV. EXACT SETS OF FIELDS

We have seen, in the previous section, how a spinor formalism may be set up to describe the formal properties of covariant derivative, space-time curvature and electromagnetic field. However, as yet, no obvious advantage has emerged from such a reformulation. Indeed, the spinor formulae just appear to be somewhat more complicated than the corresponding tensor formulae! There are, however, certain results which can be obtained even at this stage which are greatly facilitated by the use of the spinor formalism. These are centred mainly on the structure of the Weyl spinor Ψ_{ABCD}, which in empty space-time describes the entire curvature, and in any case may be thought of as defining the free gravitational field. In particular, the classification scheme for Ψ_{ABCD} is much more easily obtained than the equivalent one for $C_{\mu\nu\rho\sigma}$ and rests simply on the fact that we can write[41]

$$\Psi_{ABCD} = \alpha_{(A}\beta_B\gamma_C\delta_{D)}.$$

Here α_A, β_A, γ_A, δ_A, are uniquely defined up to proportionality and determine four gravitational principal null directions. As another example, the properties of the Bel-Robinson tensor, whose spinor equivalent is $\Psi_{ABCD}\overline{\Psi}_{P'Q'R'S'}$, become obvious in the spinor formalism. These matters will not be entered into here, however, and the interested reader is refered to the literature[41].

Instead, I shall concentrate on setting up a general framework for the discussion of sets of interacting fields in flat or curved background space-times, or in general relativity itself. The gravitational field is then described by the spinor Ψ_{ABCD} which, for the pruposes of the general discussion, can be treated on a similar footing to the other fields under consideration. The key concept will be that of an exact set of interacting fields.[44] Once one has an exact set, one is ensured that the fields will propagate correctly through

space-time; in the case of general relativity, they propagate correctly and simultaneously generate the structure of the space-time. The appropriate form of initial value problem, discussed in one aspect in section V, is one based on characteristic initial hypersurfaces. For an exact set, the initial data will be complete and irrededundant (without constraints) so that the counting of degrees of freedom becomes a simple matter. The simplifications and unifications that are obtained are a direct result of the consistent use of two-component spinors. A corresponding tensor treatment would, on the other hand, be exceedingly complicated.

Let us consider a system of fields

$$(4.1) \qquad \psi_{A\,B\,\ldots\,G}^{\;P'Q'\,\ldots\,U'}, \quad \phi_{A\,B\,\ldots\,H}^{\;P'Q'\,\ldots\,S'}, \quad \chi_{A\,B\,\ldots\,E}^{\;P'Q'\,\ldots\,T'},$$

where each spinor is to be symmetric in all its unprimed indices and symmetric in all its primed indices -- that is to say, it is to be irreducible under the group of all homogeneous (and proper orthochronous) local Lorentz transformations. Either, or both, sets of indices may be vacuous. (The reason for writing all the primed indices here in a contravariant form and the unprimed ones in a covariant form is simply notational convenience for what follows.) It is, in fact, a well-known result that any spinor whatever can be represented in terms of ($\in$'s and) spinors which, like (4.1), are symmetric in all primed and all unprimed indices. (For an example, see (3.24) or (3.35).) Any set of interacting (finite-dimensional) locally Lorentz covariant fields can be represented as a set (4.1).

Suppose the fields (4.1) are subject to a set of covariant differential equations involving the operator $\nabla_{A}^{B'}$. Then (4.1) will be called an exact set of fields if, at each event O, the following two conditions are satisfied:

(a) all the symmetrized derivatives

$$(4.2a) \quad \ldots, \; \nabla_{(J\,\ldots}^{(V'\cdots}\nabla_{L}^{X'}\psi_{A\,\ldots\,H)}^{P'\,\ldots\,U')}, \; \ldots, \; \nabla_{(J\,\ldots}^{(V'\cdots}\nabla_{M}^{Y'}\chi_{A\,B\,\ldots\,E)}^{P'Q'\,\ldots\,T')}, \; \ldots$$

(including the "zero times" differentiated fields (4.1)) are <u>independent</u> (i.e. they can take independently arbitrary values at 0) and

(b) all the unsymmetrized derivatives

$$(4.2b)\quad \ldots,\ \nabla_{K\ldots}^{W'\ldots}\nabla_{N'}^{Z'}\psi_{A\ \ldots H}^{P'\ \ldots\ U'},\ \ldots,\ \nabla_{J'\ \ldots}^{V'\ \cdots}\nabla_{N'}^{Z'}\chi_{A\ B\ \ldots\ E}^{P'Q'\ \ldots\ T'},\ \ldots$$

are <u>determined</u>, at 0, by the values of the symmetrized derivatives (4.2a) at 0, by virtue of the differential relations satisfied by the fields.

An exact set (4.1) will be called <u>invariant</u> if the expressions for (4.2b) in terms of (4.2a) are the same whichever event 0 is chosen and are locally Lorentz covariant (i.e. (4.2b) are expressed as definite spinorial combinations of ϵ's, and the spinors (4.2a), involving no extraneous quantities other than scalar constants).

As a simple example of an invariant exact set, consider a Maxwell field in charge-free Minkowski space-time. We may, according to (3.35), express the field in terms of a single symmetric spinor ϕ_{AB}. The source-free Maxwell equations are then[45]

$$(4.3)\qquad \nabla^{AP'}\phi_{AB} = 0.$$

By (3.14), this equation states the symmetry of $\nabla_C^{P'}\phi_{AB}$ in CA. In addition, the symmetry in AB now implies $\nabla_C^{P'}\phi_{AB} = \nabla_{(C}^{P'}\phi_{AB)}$. Next consider $\nabla_D^{Q'}\nabla_C^{P'}\phi_{AB}$. The operators $\nabla_D^{Q'}$, $\nabla_C^{P'}$ here commute, so that we have symmetry in DAB as well as in CAB. Thus $\nabla_D^{Q'}\nabla_C^{P'}\phi_{AB} = \nabla_{(D}^{Q'}\nabla_C^{P'}\phi_{AB)}$. Symmetry in Q'P' then also follows from the commuting of the operators. Repeating the argument with higher derivatives we get generally,

$$(4.4)\qquad \nabla_C^{P'}\nabla_D^{Q'}\cdots\nabla_F^{S'}\phi_{AB} = \nabla_{(C}^{(P'}\nabla_D^{Q'}\cdots\nabla_F^{S')}\phi_{AB)}.$$

Thus, condition (b) for exactness is (trivially) satisfied, as is the condition for invariance. Since (4.3) is linear and derivatives commute, all algebraic relations satisfied by the derivatives (4.4) at 0 must be linear. Such relations would necessarily emerge as linear operations on the indices of (4.4) because of the invariance of (4.3). But (4.4) expresses complete symmetry, so no further relations can in fact emerge. Thus condition (a) for exactness is also satisfied, so that ϕ_{AB} forms, by itself, an invariant exact set.

The analogous argument applies, in fact, to a zero rest-mass free field of any spin s (> 0) in Minkowski space-time. For (4.3) generalizes to

$$\nabla^{AP'} \phi_{AB\ldots K} = 0 \tag{4.5}$$

where $\phi_{AB\ldots K}$ is symmetric with 2s indices, whence it follows that

$$\nabla_{L}^{P'} \cdots \nabla_{N}^{R'} \phi_{AB\ldots K} = \nabla_{(L}^{(P'} \cdots \nabla_{N}^{R')} \phi_{AB\ldots K)}. \tag{4.6}$$

Thus $\phi_{AB\ldots K}$ forms an invariant exact set. In addition to the Maxwell field (s=1) this incorporates the Weyl neutrino field ($s=\frac{1}{2}$) and the linearized Einstein gravitational field (s=2). The case s=0 is also essentially the same, but in place of (4.5) we must have the second order (D'Alembert) equation

$$\nabla^{AP'} \nabla_{AP'} \phi = 0. \tag{4.7}$$

It still follows [46] that

$$\nabla_{L}^{P'} \cdots \nabla_{N}^{R'} \phi = \nabla_{(L}^{(P'} \cdots \nabla_{N)}^{R')} \phi \tag{4.8}$$

and ϕ forms an invariant exact set.

Consider, now, a Maxwell field with a specified source distribution. In place of (4.3) we have

(4.9) $$\nabla^{AP'} \phi_{AB} = \gamma_{B}^{P'}$$

where $\gamma_{BP'} = \overline{\gamma}_{P'B}$ represents the given charge-current vector subject to the divergence condition

(4.10) $$\nabla^{AB'} \gamma_{AB'} = 0.$$

Instead of (4.4) we now get

(4.11a) $$\nabla_{C}^{P'} \phi_{AB} = \nabla_{(C}^{P'} \phi_{AB)} - \tfrac{2}{3} \in_{C(A} \gamma_{B)}^{P'} ,$$

(4.11b) $$\nabla_{C}^{P'} \nabla_{D}^{Q'} \phi_{AB} = \nabla_{(C}^{(P'} \nabla_{D}^{Q')} \phi_{AB)} - \tfrac{1}{2} \in_{CD} \in^{P'Q'} \nabla_{(A}^{R'} \gamma_{B)R'} -$$

$$\tfrac{5}{6} \nabla_{(C}^{(P'} \{ \in_{D)(A} \gamma_{B)}^{Q')} \} + \tfrac{1}{6} \nabla_{(A}^{(P'} \{ \in_{B)(C} \gamma_{D)}^{Q')} \}$$

and so on for higher derivatives: the unsymmetrized n^{th} derivative of ϕ_{AB} differs from the symmetrized n^{th} derivative by an expression linear in the $(n-1)^{th}$ derivative of $\gamma_{A}^{P'}$. This, and the consequent fact that ϕ_{AB} forms an exact set, follows by the same argument as in the source-free case except for the essential modification entailed by the replacing of (4.3) by (4.9); i.e. when a pair of indices is interchanged, a term involving $\gamma_{A}^{P'}$ may be introduced.

However, it is clear that ϕ_{AB} by itself does not form an <u>invariant</u> exact set here, since the extraneous quantity $\gamma_{A}^{P'}$ appears in (4.11). On the other hand, we may admit $\gamma_{A}^{P'}$ as a variable field; but then ϕ_{AB} and $\gamma_{A}^{P'}$ together do not, as things stand, form an exact set. For (4.10) is concerned only with the part of $\nabla_{B}^{Q'} \gamma_{A}^{P'}$ which is skew both in Q'P' and BA. The part which is symmetric in BA and skew in Q'P', for example, will remain undetermined. Further equations would need to be imposed on $\gamma_{A}^{P'}$ to give an exact set. This could be done in many ways, but of most interest are the cases when $\gamma_{A}^{P'}$ is given

by the charge-current vector of a physical field (or fields), say a Dirac field or a Schrödinger-Klein-Gordon field.

Consider, first, the Dirac case. In the two-component spinor form,[40,42] the Dirac field is represented as a pair of spinors ψ_A, $\chi^{P'}$ subject to

$$\nabla^{AP'} \psi_A = \mu \chi^{P'}, \quad \nabla_{AP'} \chi^{P'} = -\mu \psi_A \tag{4.12}$$

where μ is a real constant defining the mass. In the absence of electromagnetism we can assume that the ∇'s commute ($\nabla_A^{P'} \equiv \partial_A^{P'}$). It is then not hard to verify that ψ_A and $\chi^{P'}$ together form an invariant exact set. Moreover, if we introduce electromagnetism via (4.9) with $\gamma_A^{P'} = k (\psi_A \overline{\psi}^{P'} + \overline{\chi}_A \chi^{P'})$, i.e.

$$\nabla^{AP'} \phi_{AB} = k (\psi_B \overline{\psi}^{P'} + \overline{\chi}_B \chi^{P'}) \tag{4.13}$$

where k is a real constant, and use the definition (3.3) for $\nabla_A^{P'}$, each of ψ_A, $\chi^{P'}$ having the same charge e, then ψ_A, $\chi^{P'}$ and ϕ_{AB} together form an invariant exact set. This may be demonstrated in essentially the same way as for the first case, except that the interchanging of derivatives may give rise to extra terms involving ϕ_{AB}, $\overline{\phi}^{P'Q'}$ (see (3.37)). For example[47]

$$\nabla_{\mathcal{B}}^{P'} \psi_A = \nabla_{(B}^{P'} \psi_{A)} - \tfrac{1}{2} \epsilon_{BA} \chi^{P'}, \tag{4.14a}$$

$$\nabla_B^{P'} \nabla_C^{Q'} \psi_A = \nabla_{(B}^{(P'} \nabla_{C}^{Q')} \psi_{A)} + \tfrac{2}{3} i e \overline{\phi}^{P'Q'} \epsilon_{B(C} \psi_{A)} \tag{4.14b}$$

$$+ i e \epsilon^{P'Q'} \{ \phi_{B(C} \psi_{A)} - \tfrac{1}{2} \psi_B \phi_{CA} \}$$

$$+ \tfrac{1}{2} \mu^2 \epsilon^{P'Q'} \epsilon_{BC} \psi_A + \tfrac{2}{3} \mu \epsilon_{A(C} \nabla_{B)}^{(P'} \chi^{Q')}.$$

The case of a Schrödinger-Klein-Gordon field Θ is similar, but in place of (4.12), (4.13) we have

$$(4.15) \qquad (\nabla^{AB'}\nabla_{AB'} - m^2)\Theta = 0; \quad \nabla^{AP'}\phi_{AB} = k\,(i\Theta\nabla_B^{P'}\bar{\Theta} - i\bar{\Theta}\nabla_B^{P'}\Theta).$$

Again we get an invariant exact set for Θ, ϕ_{AB}.

There is an alternative method of dealing with charged fields which involves introducing the electromagnetic potential $\alpha_A^{P'}$ explicitly and treating $\alpha_A^{P'}$ as a new field. In this approach, $\nabla_A^{P'}$ is not defined as in (3.3) but simply put equal to the gradient operator $\partial_A^{P'}$. Then the ∇'s commute, but $\alpha_A^{P'}$ now enters explicitly in (4.12), or in the first of equations (4.15), with $\nabla_A^{P'}$ replaced by $\nabla_A^{P'} - i\,e\,\alpha_A^{P'}$. In order to get an exact set we must impose a restriction on $\alpha_A^{P'}$, such as the Lorenz condition: $\nabla_{P'}^{A}\alpha_A^{P'} = 0$. Then (3.36) becomes $\phi_{AB} = -\nabla_{AC'}\alpha_B^{C'}$. We also have $\bar{\alpha}_{A'B} = \alpha_{BA'}$ since α_μ is real, whence $\bar{\phi}^{P'Q'} = \nabla^{AP'}\alpha_A^{Q'}$. Using these equations, it is not hard to show that ψ_A, $\chi^{P'}$, ϕ_{AB}, $\alpha_A^{P'}$ and Θ, ϕ_{AB}, $\alpha_A^{P'}$ each form an invariant exact set.

This alternative approach is, perhaps, conceptually a little simpler than the one using (4.12) and the consequent non-commutativity of the ∇'s and it is useful in some contexts. However, it is more in keeping with the philosophy being adopted here not to introduce gauge dependent quantities, such as α_μ explicitly into the formalism. In the case of electromagnetism the gauge independent approach in fact involves somewhat simpler formulae (e.g. (4.14)) than does the approach explicitly involving α_μ. Moreover the theories which operate in curved space-time can be treated here only by virtue of the existence of a gauge independent (i.e. coordinate independent) treatment.

Consider, first, the case of a set of fields in a <u>given</u> Riemannian background space-time. The curvature quantities Ψ_{ABCD}, $\Phi_{ABC'D'}$ and Λ (see (3.24)) are then given at each event 0. These quantities enter into the commutation relations for the ∇'s (cf. (3.29), (3.30)). Certain flat space-time exact sets can be transcribed for curved space-time by simply adding the appropriate terms (cf. (3.38)) to deal with interchanges of indices of ∇'s, this resulting in an exact set. For example, in the case of a Maxwell field, in place of (4.4) we have[47]

$$(4.16)\qquad \nabla_C^{P'}\nabla_D^{Q'}\phi_{AB} = \nabla_{(C}^{(P'}\nabla_{D)}^{Q')}\phi_{AB} + \epsilon^{P'Q'}\left(\tfrac{1}{2}\Psi_{DAB}{}^{X}\phi_{CX} - \tfrac{3}{2}\Psi_{C(AB}{}^{X}\phi_{D)X} + 3\Lambda\,\epsilon_{C(D}\phi_{AB)}\right) - \tfrac{3}{2}\epsilon_{C(D}\Phi_{A}{}^{XP'Q'}\phi_{B)X}$$

etc. and ϕ_{AB} gives an exact set. It is clearly not an invariant exact set, however. Only in the case of de Sitter space-time (Λ = const., $\Psi_{ABCD} = 0$, $\Phi_{AB}^{P'Q'} = 0$) does an invariant exact set in fact arise. The Maxwell-Dirac equations can also be transcribed in the same way for curved space time and an exact set is obtained. On the other hand, the zero rest-mass equation (4.5) for spin $s \geqslant \frac{3}{2}$ does <u>not</u>, as it stands, lead to an exact set for $\phi_{AB\ldots K}$ for a given space-time which is <u>not conformally flat</u>. For by (4.5), (3.30a):

$$(4.17)\qquad 0 = \nabla_{P'}^{B}\nabla^{AP'}\phi_{ABCD\ldots K} = \nabla_{P'}^{(B}\nabla^{A)P'}\phi_{ABCD\ldots K}$$

$$= (2-2s)\,\Psi^{ABL}{}_{(C}\phi_{D\ldots K)ABL}$$

so that the $\phi_{AB\ldots K}$ are not independent unless $\Psi_{ABCD} = 0$ (conformal flatness condition). In fact there are essential difficulties for zero rest-mass fields of spin $s \geqslant \frac{3}{2}$ in conformally curved background space-times[48].

For general relativity proper, the situation is somewhat different. Here, curvature quantities are to be considered as field variables. The degrees of freedom of the gravitational field can be described by the conformal spinor Ψ_{ABCD} (cf. (3.27)). The quantities $\Phi_{A\ B}^{P'Q'}$ and Λ are defined directly in terms of the remaining fields -- and possibly a cosmological constant λ -- through their energy-momentum tensor $T_{\mu\nu}$:

$$(4.18) \qquad \kappa T_{AB'CD'} = \Phi_{ACB'D'} + (3\Lambda - \tfrac{1}{2}\lambda)\in_{AC}\in_{B'D'},$$

κ being a positive constant. In particular, the Einstein-Maxwell equations (with cosmological constant λ) are given by

$$(4.19) \qquad \Phi_{A\ B}^{C'D'} = \kappa\,\phi_{AB}\,\bar{\phi}^{C'D'}, \quad \Lambda = \tfrac{1}{6}\lambda\ .$$

The Bianchi identity (3.31) becomes

$$(4.20) \qquad \nabla^{AP'}\Psi_{ABCD} = -\kappa\,\bar{\phi}^{P'Q'}\nabla_{BQ'}\phi_{CD},$$

which, with (4.3) and the commutator equations (3.30), has been shown[41] to lead to an invariant exact set for Ψ_{ABCD}, ϕ_{AB}. Also the Einstein empty-space equations (put $\phi_{AB} = 0$) give an invariant exact set for Ψ_{ABCD} alone. (Note that, in this case (4.20) becomes the free-field spin 2 zero rest-mass equation (4.5) with $\Psi_{ABCD} = \phi_{ABCD}$. No trouble here arises from (4.17) since the last term now vanishes by symmetry considerations.) There are also many other sources for the gravitational field which, with Ψ_{ABCD} give an invariant exact set.

Finally, some sets of fields which, by themselves, do not give exact sets, can be completed to exact sets by including certain contracted derivatives of the fields as additional fields. A trivial example is the Fock-Feynman-Gell-Mann equation for a spin $\frac{1}{2}$ particle.[49] The single two-component spinor ψ_A is

subject to $\nabla_{B\,P'}\,\nabla^{AP'}\,\psi_A = -\mu^2\,\psi_B$ and, by itself (or with ϕ_{AB}) does not form an exact set. But, if we include $\chi^{P'} = \mu^{-1}\,\nabla^{AP'}\,\psi_A$ as an additional field, we are back with the Dirac equation (4.12), so ψ_A, $\chi^{P'}$ (and ϕ_{AB}) give an invariant exact set. Somewhat similar is the case of a Dirac-Fierz particle[50] of spin $s \geqslant 1$, in flat space-time. Suppose electromagnetism to be absent and write

$$(4.21)\qquad \nabla^{AP'}\,\psi_{AB\ldots\ D}^{\ Q'\ldots\ T'} = \mu\,\chi_{B\ \ldots\ D}^{\ P'Q'\ldots\ T'},\quad \nabla_{AP'}\,\chi_{B\ldots D}^{\ P'Q'\ldots T'} = -\mu\psi_{AB\ldots\ D}^{\ Q'\ldots\ T'}$$

where $\psi_{A\ \ldots D}^{\ Q'\ldots T'}$, $\chi_{B\ \ldots D}^{\ P'\ldots T'}$ have each a total of 2s indices and are symmetric in both sets of indices. We can introduce 2s-1 new spinors each with 2s indices and similarly symmetric, but with differing numbers of primed and unprimed indices: in their natural order these 2s+1 spinors form a linear sequence of spinors each of which is obtainable from its immediate neighbours by a differentiation contracted on one index. Together, these spinors form an invariant exact set. For example, if s=1 we have the invariant exact set ψ_{AB}, $\chi_B^{P'}$, $\zeta^{P'Q'}$ with

$$(4.22)\qquad \nabla^{AP'}\,\psi_{AB} = \mu\,\chi_B^{P'},\quad \nabla^{AP'}\,\chi_A^{Q'} = \mu\,\zeta^{P'Q'};\quad \nabla_{AP'}\,\zeta^{P'Q'} = -\mu\chi_{A,}^{Q'}$$

$$\nabla_{AP'}\,\chi_B^{P'} = -\mu\psi_{AB}\ .$$

However, if electromagnetism (with $s \geqslant 1$) or gravity (with $s \geqslant \frac{3}{2}$) are to be incorporated the situation becomes more complicated, since as it stands, (4.21) is inconsistent with (3.37) (when $s \geqslant 1$)[51] or with (3.30) (when $s \geqslant \frac{3}{2}$).[52]

V. INITIAL DATA ON A NULL CONE

In the previous section the most important of the explicitly known physical fields, gravitation included, were exhibited in the form of exact sets. Interactions between fields found expression either through "field equations" such as (4.12), 4.13), (4.20), or through "commutator equations" for the ∇'s (cf. (3.30), (3.37), (3.38)). (Note that in this formalism, the Bianchi identity counts as a "field equation"!) The entire formalism is expressed in terms of field quantities rather than gauge dependent potentials. Thus, the electromagnetic potential α_μ never appears explicitly, nor is there any place for the explicit appearance of gravitational potentials (i.e. expressions of the $g_{\mu\nu}$ in terms of coordinates). In the present section this "geometrical attitude" is exploited further. Assuming analyticity of all the quantities involved, if all the derivatives (4.2b) are known at one event 0, then by means of power series, we can "walk" from event to event, thus calculating the fields (4.1) and their derivatives (4.2b) at every other event in the space-time. This gives a method of exploring the space-time in a way which is, in principle, completely coordinate-free.[53] It is not, however, apparently a very practical method (except in certain special cases),[41] although it is possible that developments in formal technique might lead to a more manageable procedure.

On the other hand, the method does lead to one important deduction concerning the nature of the exact set condition, and it is this which I wish to present here. Roughly speaking, in the same way that the derivatives (4.2b) enable us to "walk" about the space-time, it is the symmetrized derivatives (4.2a) which enable us to "walk" up the null cone of the event 0 with that component of each field which is associated with the null direction in the cone.[44] In consequence, condition (a) for exactness will tell us that, for each field, this component can be specified as an essentially arbitrary

function on the null cone of 0, while condition (b) tells us that knowledge of this function defines the fields <u>everywhere</u> [53] throughout space-time. We thus have a form of initial value problem for which the data can be specified in an arbitrary (i.e. "constraint-free") way on a null cone. The argument is presented here only in its simplest form, where it is assumed that everything is analytic. The results have a much wider domain of applicability than this, however, the analyticity being inessential (if due caution is exercised) to anything but the method of proof used. A full discussion would take us too far afield from the immediate purposes here, but the comparatively simple (analytic) argument I shall present should be adequate to give the general ideas and to provide some of the motivations for ideas used in sections VI and VII.

I assume, then, that the space-time $\mathcal{M}$ is an analytic manifold [4] with analytic metric ($g_{\mu\nu}$ and $\sigma_\mu{}^{BC'}$ analytic on $\mathcal{M}$), the operator ∇_μ also being analytic (i.e. yields an analytic field whenever applied to an analytic field) and satisfying

$$\nabla_\lambda g_{\mu\nu} = 0 , \quad \nabla_\mu \sigma_\nu{}^{AB'} = 0 , \quad \nabla_\mu \epsilon_{AB} = 0 , \tag{5.1a}$$

$$\nabla_\mu (\theta^{\cdots} + \chi^{\cdots}) = \nabla_\mu \theta^{\cdots} + \nabla_\mu \chi^{\cdots} , \tag{5.1b}$$

$$\nabla_\mu (\theta^{\cdots} \chi^{\cdots}) = (\nabla_\mu \theta^{\cdots}) \chi^{\cdots} + \theta^{\cdots} \nabla_\mu (\chi^{\cdots}) \tag{5.1c}$$

$$\overline{\nabla_\mu \theta^{\cdots}} = \nabla_\mu \overline{\theta^{\cdots}} . \tag{5.1d}$$

This includes the possibility of ∇_μ being the modification of covariant derivative which incorporates the electromagnetic potential when applied to charged fields. It also includes the possibility of torsion. Since I shall not be concerned with torsion or further possibilities here, the ∇_μ will be further pinned down by the condition that if θ is any <u>uncharged scalar</u> field,

$$\nabla_\mu \nabla_\nu \theta = \nabla_\nu \nabla_\mu \theta \tag{5.1e}$$

$$\nabla_\mu \theta = \partial_\mu \theta \tag{5.1f}$$

where ∂_μ is the ordinary gradient operator.

Let c be a smooth curve in $\mathcal{M}$ and let ℓ^μ be a tangent vector defined at each event of c. Suppose for some $\Theta^{\cdots}_{\cdots}$,

$$\ell^\mu \nabla_\mu \Theta^{\cdots}_{\cdots} = 0 \tag{5.2}$$

at each event of c. Then we say that $\Theta^{\cdots}_{\cdots}$ is constant along c (with respect to ∇_μ) and write

$$[\Theta^{\cdots}_{\cdots}]_P \overset{c}{=} [\Theta^{\cdots}_{\cdots}]_Q \tag{5.3}$$

for $\Theta^{\cdots}_{\cdots}$ evaluated at any two events P, Q along c. Unless the commutators of the ∇'s vanish, the notion of equality defined by (5.3) generally depends on the choice of curve c which connects the events in question. (If c is a closed curve originating and terminating at the same event P(=Q), then (5.3) would not generally imply that the original and final fields at P are equal in the ordinary sense. For an infinitesimal loop, the discrepancy comes directly from the ∇ commutator term.)

If ℓ^μ itself can be chosen constant along c, i.e.

$$\ell^\nu \nabla_\nu \ell^\mu = 0, \tag{5.4}$$

then c is a geodesic in $\mathcal{M}$. Let r be a real parameter on the geodesic c, scaled so that

$$\ell^\mu \nabla_\mu r = 1. \tag{5.5}$$

(Thus r will be an affine parameter on c. ℓ^μ and r are "uncharged.") Let the event 0 on c be given by r=0. To represent an event X on c, define a vector x^μ at 0 by

$$x^\mu = r\ell^\mu \tag{5.6}$$

where r is the parameter value at X. (It follows at once from (5.4), (5.1) and (5.5) that (5.6) is independent of the scaling of ℓ^μ. If $\ell^\mu \to \lambda \ell^\mu$, then λ must be constant along c to preserve (5.4), whence $r \to \lambda^{-1} r$.) Thus we may say that x^μ gives the position vector of X relative to 0. As c varies over all geodesics through 0, the components of x^μ give, in fact, the system of normal coordinates with origin 0. That is to say, the direction of x^μ is that of the geodesic c through X and the length (or affine "extent" if x^μ is null) of x^μ is the same as that of c from 0 to X. (This follows directly from (5.1), (5.4), (5.5), (5.6).)

Let $\phi^{\cdots}_{\cdots}$ be analytic at 0, then for X not too far from 0 on the geodesic c we have

$$(5.7)\qquad [\phi^{\cdots}_{\cdots}]_X \triangleq [\phi^{\cdots}_{\cdots}]_0 + \frac{x^\mu}{1!}[\nabla_\mu \phi^{\cdots}_{\cdots}]_0 + \frac{x^\mu x^\nu}{2!}[\nabla_\mu \nabla_\nu \phi^{\cdots}_{\cdots}]_0 + \ldots .$$

To see this,[54] we note that if r' is the parameter assigned to a variable point X' between 0 and X on c, the quantity

$$(5.8)\qquad \phi^{\cdots}_{\cdots} + \frac{(r-r')\ell^\mu}{1!}\nabla_\mu \phi^{\cdots}_{\cdots} + \frac{(r-r')^2 \ell^\mu \ell^\nu}{2!}\nabla_\mu \nabla_\nu \phi^{\cdots}_{\cdots} + \ldots .$$

defined at X' is constant along c by (5.2), (5.4), (5.5) (r being constant and r' satisfying $\ell^\mu \nabla_\mu r' = 1$). The convergence and appropriate uniformity of (5.8), for X near enough to 0, is implicit in the analyticity of the quantities involved. When r' = 0 and r' = r, we get, respectively, the right- and left-hand sides of (5.7).

Relation (5.7) tells us how knowledge of all the unsymmetrized derivatives (4.2b) of a set of fields at 0 can be employed in the determination of the fields at all other events (not too far from 0 in the first instance, but "distant" events can be reached using several steps).[53] Let us now examine

the role of the symmetrized derivatives (4.2a). Consider the events X which lie on the null cone $\mathcal{N}$ of O. Then c will be a null geodesic and ℓ^μ a null vector, whose spinor equivalent is thus of the form

$$\ell^{AB'} = \xi^A \bar{\xi}^{B'} \tag{5.9}$$

(choosing ℓ^μ as future-pointing). We can make a single choice of ξ^A for each null geodesic c, defining ξ^A as constant and uncharged along c. Multiply (5.7) (for a spinor $\phi_{AB\ldots E}^{P'Q'\ \ S'}$) by $\xi^A \xi^B \ldots \xi^E \bar{\xi}_{P'} \ldots \bar{\xi}_{S'}$ (where, because of the constancy of ξ^A along c, we can unambiguously write the ξ's outside the brackets):

$$\begin{aligned}
&\xi^A \ldots \xi^E \bar{\xi}_{P'} \ldots \bar{\xi}_{S'} \left[\phi_{A\ldots E}^{P'\ldots S'}\right]_X \\
&\overset{c}{=} \xi^A \ldots \xi^E \bar{\xi}_{P'} \ldots \bar{\xi}_{S'} \left[\phi_{A\ldots E}^{P'\ldots S'}\right]_0 \\
&- \tfrac{r}{1!}\, \xi^A \ldots \xi^E \xi^F \bar{\xi}_{P'} \ldots \bar{\xi}_{S'} \bar{\xi}_{T'} \left[\nabla_F^{T'} \phi_{A\ldots E}^{P'\ldots S'}\right]_0 \\
&+ \tfrac{r^2}{2!}\, \xi^A \ldots \xi^E \xi^F \xi^G \bar{\xi}_{P'} \ldots \bar{\xi}_{S'} \bar{\xi}_{T'} \bar{\xi}_{U'} \left[\nabla_F^{T'} \nabla_G^{U'} \phi_{A\ldots E}^{P'\ldots S'}\right]_0 - \ldots
\end{aligned} \tag{5.10}$$

Because of the symmetry of $\xi^A \ldots \xi^H \bar{\xi}_{P'} \ldots \bar{\xi}_{V'}$, it is really the symmetrized derivatives $\left[\nabla_{(F\ldots}^{(T'\cdots} \nabla_H^{V'} \phi_{A\ldots E)}^{P'\ldots S')}\right]_0$ that are relevant here. Knowledge of all these symmetrized derivatives at O will determine the quantity

$$\phi = \xi^A \ldots \xi^E \bar{\xi}^{P'} \ldots \bar{\xi}^{S'} \left[\phi_{A\ldots EP'\ldots S'}\right]_X \tag{5.11}$$

at each point of $\mathcal{N}$.

The complex number ϕ will be called the null-datum for the field $\phi_{\ldots}^{\ldots}$ at the point X of $\mathcal{N}$. Since we have made a definite choice of ξ^A for

each null direction at 0, we may regard ϕ as a scalar function defined on the null cone of 0, this function being determined by the symmetrized derivatives of $\phi_{\cdots}^{\cdots}$ at 0. Conversely, the values of ϕ on $\mathcal{N}$ determine these symmetrized derivatives. For the null cone $\mathcal{N}$ may be parameterized by the real number r (r = 0 giving the vertex) and by the complex ratio $\xi^0 : \xi^1$ at 0. The coefficient of $(-r)^n$ in (5.10) is

$$(5.12) \qquad \xi^A \cdots \xi^E \xi^F \cdots \xi^H \bar{\xi}_{P'} \cdots \bar{\xi}_{S'} \bar{\xi}_{T'} \cdots \bar{\xi}_{V'}$$
$$\times \left[\nabla_F^{T'} \cdots \nabla_H^{V'} \phi_{A \cdots E}^{P' \cdots S'} \right]_0$$

which, for each value of the ratio $\xi^0 : \xi^1$, will therefore be determined by ϕ. Hence, (5.12) will be determined by ϕ as a function of ξ^A, $\bar{\xi}_{P'}$. But ξ^A and $\bar{\xi}_{P'}$ are algebraically independent. Therefore (5.12), being a polynomial in ξ^0, ξ^1, $\bar{\xi}_{0'}$, $\bar{\xi}_{1'}$, will determine its coefficients $\left[\nabla_{(F \cdots}^{(T'} \nabla_H^{V'} \phi_{A \cdots E)}^{P' \cdots S')} \right]_0$ uniquely.

The significance of the concept of an exact set should now be apparent. Condition (b) for exactness for a set of fields $\psi_{\cdots}^{\cdots}, \ldots, \chi_{\cdots}^{\cdots}$ ensures that the values of their null-data $\psi, \ldots, \chi$, on the null cone $\mathcal{N}$, determine the fields throughout the space-time $\mathcal{M}$. Condition (a) for exactness ensures that the various null data can be chosen freely on $\mathcal{N}$, that is, there are no constraints. For any exact set, therefore, the null-data are a complete irredundant set of initial data on any null cone.

We are now in a position to count the number of degrees of freedom for the fields. We must bear in mind that $\mathcal{N}$, being a characteristic hypersurface will require just one-half as many real numbers per point of the hypersurface as would be required in the case of a spacelike hypersurface.[55] In the present case, we shall have just one complex number (i.e. two real numbers) per

point of $\mathcal{N}$, for each of the fields (4.1) belonging to the exact set. This will hold unless the field is specified as being Hermitian (in which case it must have an equal number of primed and unprimed indices). For a Hermitian field the null-datum must be real. Thus, the number of "degrees of freedom" for the fields, being defined as one-half the number of real variables required for an initial spacelike hypersurface, is twice the number of non-Hermitian fields occurring in (4.1) plus the number of Hermitian fields.

Let us just briefly check this with some known results. For the case of Maxwell's free-space equations, in a given background space-time, we have just one field ϕ_{AB}, so the number of degrees of freedom is two. (In the usual formalism, this is one-half the number obtained from: six degrees of freedom for $\underline{E}$ and $\underline{H}$ minus two for the constraints div $\underline{E}$ = 0, div $\underline{H}$ = 0.) In the case of a Dirac field we have two spinors ψ_A, $\chi^{A'}$ in (4.1). Thus we have four degrees of freedom. (In the usual formalism one thinks of these as given by the four complex components of a single Dirac 4-spinor.) For a real scalar (Schrödinger-Klein-Gordon or wave) field we have one "Hermitian" field and thus, just one degree of freedom. (In the usual formalism, the initial data is the field together with its time derivative.) For the Dirac-Fierz higher spin fields (4.21), we require a total of 2s+1 non-Hermitian spinors (cf. (4.22)) in order to get an exact set, so there are 2s+1 degrees of freedom. (In the usual formalism this comes about in a different way [56].) Finally, in the case of general relativity, we see at once that there are just two degrees of freedom for the gravitational field, since this is defined by the single non-Hermitian spinor Ψ_{ABCD}. (In the usual formalism the result is the same, but is not nearly so directly obtained, there being many redundant variables and constraints to be taken into account [57].)

I shall not enter into any of the difficult questions concerned, say, with removing the condition of analyticity, or of replacing the null cone $\mathcal{N}$ by a more general characteristic hypersurface. We should expect, for example, the null-data on the future half of $\mathcal{N}$ (or, say, on $C_+(0)$, cf (2.14a)) to define the fields in $J_+(0)$ (at least at those events of $J_+(0)$ not too far from 0, see (2.7a) and section VIII) in a "reasonable" way, etc. But such questions as domains of dependence and stability, etc.,[55] cannot be answered by considerations only of analytic fields. The only question of detail I shall examine here concerns the nature of the null-data at the vertex 0 of $\mathcal{N}$. Since 0 is a singular point of $\mathcal{N}$, it is not immediately clear what kind of smoothness conditions the null-data should satisfy there.

In this connection, the dependence of ϕ on ξ^A should be kept in mind. For under

$$\xi^A \longrightarrow \alpha\xi^A \tag{5.13}$$

we have

$$\phi \longrightarrow \alpha^p\,\bar{\alpha}^q\,\phi \tag{5.14}$$

where $\phi_{\cdots}^{\cdots}$ has, respectively p unprimed and q primed indices. (At any point X of $\mathcal{N}$, ξ^A is, of course, defined up to proportionality as it corresponds to the tangent to the generator of $\mathcal{N}$ through X). Generally, a quantity defined on $\mathcal{N}$ which transforms as ϕ in (5.14) under (5.13) will be said to have <u>weights</u> (p, q). (Note, in particular, that r has weights (1, 1).) The integer on half-odd integer

$$s = \frac{p - q}{2} \tag{5.15}$$

will also be referred to as the <u>spin-weight</u> of ϕ.

Now, let us consider a scalar quantity ϕ, of weights (p, q), defined on $\mathcal{N}$, but without reference to the whole space-time $\mathcal{M}$. We may take for the definition of analyticity for ϕ, simply that a series of the form

$$(5.16)\qquad [\phi]_X \overset{c}{=} \xi^A \dots \xi^E \bar{\xi}_{P'} \dots \bar{\xi}_{S'} \sum_{n=0}^{\infty} x^{\alpha_1} x^{\alpha_2} \dots x^{\alpha_n} \times \phi_{(n)\,A \dots E\,\alpha_1 \alpha_2 \dots \alpha_n}^{P' \dots S'}$$

should exist giving ϕ on $\mathcal{N}$, convergent in some neighbourhood of 0 (for x^α null: $x^{AB'} = r\,\xi^A \bar{\xi}^{B'}$), the $\phi_{(n)\dots}^{\dots}$ being constants defined at 0. The $\xi^A, \dots, \xi^E$ are p in number and the $\bar{\xi}_{P'}, \dots, \bar{\xi}_{S'}$ are q in number, giving ϕ of weights (p, q) as required. Clearly the ϕ defined by (5.11), (5.10) will, in this sense, be analytic. (Since ϕ is a scalar, the relevance of the "c" in (5.16) is only to charged fields. But we can ignore it in any case if we choose the electromagnetic potential to be analytic -- again in the sense of (5.16).)

Let us consider the nature of the regularity of ϕ at the vertex 0, implied by (5.16). If we normalize ξ^A according to $|\xi^0|^2 + |\xi^1|^2 = 1$, then r, Θ, φ can be used as spherical polar coordinates for $\mathcal{N}$, where $e^{i\varphi} \tan \frac{1}{2}\Theta = \xi^0 / \xi^1$. Positive values of r describe the future half of $\mathcal{N}$ and negative values of r, the past half. By rotating the null-flag planes, we can further specialize ξ^A so that

$$(5.17)\qquad \xi^0 = e^{\frac{1}{2} i\varphi} \sin \tfrac{1}{2}\Theta\,, \quad \xi^1 = e^{-\frac{1}{2} i\varphi} \cos \tfrac{1}{2}\Theta\,.$$

For simplicity, suppose, first, that $p = q = 0$. Then it is not hard to see from (5.16) that (for small constant values of r) the function ϕ can involve, at order r^n, spherical harmonics only of orders 0, 1, ..., n. This may be compared with the behaviour of a C^n differentiable function at the origin of

coordinates in an ordinary Euclidean three-space, with ordinary spherical polar coordinates r, θ, φ. In this case, at order r^n, spherical harmonics only of orders n, n-2, n-4, ..., 1 (n odd) or ..., 0 (n even) would be involved. Now, consider the general case where p and q are arbitrary non-negative integers. Since ϕ has a spin-weight s, which is different from zero if $p \neq q$, we shall require the spin-weighted spherical harmonics that will be introduced in section VI. Assuming the relevant results of section VI (cf. (6.45), particularly), we can see that (5.16) implies that at order r^n, ϕ can involve the appropriate spin s spherical harmonics ${}_sY_{\ell,m}$ only for $\ell \leqslant \frac{1}{2}(p+q) + n$. That is,

$$(5.18) \qquad \phi = \sum_{n=0}^{\infty} \sum_{\ell=0}^{q+s+n} \sum_{m=-\ell}^{\ell} {}_sY_{\ell,m}\, \phi_{n,\ell,m}\, r^n$$

where the $\phi_{n,\ell,m}$ are constants. Thus, the required C^N smoothness condition at 0, for a null-datum, is that for small r it should resemble (5.18) to order r^N.

In the case of gravitation, the null-datum Ψ has a special significance. Choose a complex null vector m^μ, orthogonal to ℓ^μ, whose real part (spacelike and of length $2^{-1/2}$) spans, with ℓ^μ, the null-flag plane of ξ^A, i.e.

$$(5.19) \qquad \xi^A \xi^B \in^{P'Q'} = (\ell^\mu m^\nu - m^\mu \ell^\nu)\, \sigma_\mu{}^{AP'}\, \sigma_\nu{}^{BQ'}.$$

Then we can write

$$(5.20) \qquad \Psi = \xi^A \xi^B \xi^C \xi^D \Psi_{ABCD} = \ell^\mu m^\nu \ell^\rho m^\sigma C_{\mu\nu\rho\sigma} .$$

Eq. (5.20) tells us that Ψ measures the "purely astigmatic" part of the geodetic deviation [58] of the null geodesics in η. If a matter tensor $T_{\mu\nu}$ is present, we can put

$$(5.21)\qquad \Phi = \xi^A \xi^B \bar{\xi}^{C'} \bar{\xi}^{D'} \Phi_{ABC'D'} = -\tfrac{1}{2} R_{\mu\nu} \ell^\mu \ell^\nu$$
$$= \kappa T_{\mu\nu} \ell^\mu \ell^\nu$$

by (4.18). Then Φ measures the "anastigmatic" part of this geodetic deviation. For the matter fields we have been considering here, it turns out that Φ is determined directly by the null-data of these matter fields (perhaps involving a derivative in the ℓ^μ direction). In particular, for the Einstein-Maxwell equations, we have $\Phi = \kappa \phi \bar{\phi}$, by (4.19), where ϕ is the Maxwell null-datum. Thus, we see that, in general relativity, the geometry of $\mathcal{N}$ is itself directly determined by the null-data of all the various fields involved.

In the special case of the Einstein vacuum equations we can go further. The intrinsic geometry of $\mathcal{N}$ is then equivalent to the null-datum Ψ. (There is, however, a certain subtlety involved here in that the intrinsic metric of $\mathcal{N}$ has vanishing determinant, but this is not serious). Thus, we can say that for a vacuum space-time $\mathcal{M}$ (which is analytic) the geometry of the whole of $\mathcal{M}$ is determined by the intrinsic geometry of the complete null cone of any one event in $\mathcal{M}$.

VI. THE SPIN COEFFICIENT METHOD

A technique in Riemannian geometry of great value is that of introducing (non-holonomic) orthonormal frames. In four dimensions, we can refer to these simply as tetrads. The analogous idea for a spinor technique is to introduce a pair of basis spinors at each point, called a dyad. Now, the tetrad components of the covariant derivatives of the tetrads define the Ricci rotation coefficients. Similarly, we may take the dyad components of the covariant derivatives of the dyads and define spin coefficients.[59] These are twelve complex functions of position defined by the dyad system. Certain of the spin coefficients have immediate geometrical interpretations in relation to the dyad directions. The spin coefficients turn out to be particularly useful in detailed calculations. One reason for this appears to be that they are all complex, so that two pieces of information are conveniently incorporated into each spin coefficient. This gives a sufficient compactness that it is possible to write out all equations explicitly. The resulting equations turn out to be fairly systematic and manageable, if a certain simplifying choice of the dyad system is made.

I shall be concerned, here, only with one particular class of applications of the spin coefficient method. This is to a study of asymptotic properties of gravitational and electromagnetic fields in asymptotically flat space-times. The method leads to a remarkably compact set of equations, from which it is possible to deduce, very simply, the various known conservation laws for such space-times.[60] These include the laws of mass and momentum conservation, incorporating the Bondi-Sachs expression[61] for the contribution (necessarily positive) carried away by gravitational radiation. Also some newly discovered "absolute" conservation laws[62] (sixteen in all) for the Einstein-Maxwell equations are readily obtained. The method is greatly enhanced by the introduction of an angular differential operator[63] (denoted by $\eth$), whose main properties will also be described here.

It is to be hoped that a thorough and detailed analysis of asymptotically flat space-times, where both gravitational and electromagnetic radiation, and perhaps also neutrino radiation, may be incorporated, will eventually lead to results of astrophysical interest. For, in the violent processes associated with gravitational collapse it may be expected that large energy losses will occur in any or all of these three forms. A quite different direction in which such an analysis might be of value, lies in the quantization programme. A thorough understanding of the wave modes in classical gravitational theory and of the way in which waves can interact and scatter one another, would seem to be a valuable prerequisite for a proper appreciation of the quantum problem.

Let us now consider the formalism. At each event in space-time, we introduce our dyad, that is, a pair of basis spinors o^A, ι^A normalized thus:

$$(6.1) \qquad o_A \iota^A = \epsilon_{AB} o^A \iota^B = - \iota_A o^A = 1.$$

A dyad always defines an effectively equivalent null tetrad

$$(6.2) \qquad l^\mu = \sigma^\mu_{AB'} o^A \bar{o}^{B'}, \quad n^\mu = \sigma^\mu_{AB'} \iota^A \bar{\iota}^{B'}, \quad m^\mu = \sigma^\mu_{AB'} o^A \bar{\iota}^{B'}.$$

Here, l^μ and n^ν are real future-pointing null vectors and m^μ, with its complex conjugate $\bar{m}^\mu$, being complex null vectors. The vector m^μ is equivalent to a pair of real orthogonal spacelike unit vectors a^μ and b^μ, where $m^\mu = 2^{-1/2} (a^\mu + ib^\mu)$. We have the orthogonality relations

$$(6.3) \qquad l_\mu l^\mu = m_\mu m^\mu = n_\mu n^\mu = l_\mu m^\mu = n_\mu m^\mu = 0$$
$$l_\mu n^\mu = - m_\mu m^\mu = 1$$

As with the tetrads, it is convenient to have a generic symbol for both o^A and ι^A. Define $\zeta_a{}^A$, $\bar{\zeta}_a{}^{A'}$ by

$$(6.4) \qquad \zeta_0{}^A = o^A , \quad \zeta_1{}^A = \iota^A , \quad \bar{\zeta}_{0'}{}^{A'} = \bar{o}^{A'} , \quad \bar{\zeta}_{1'}{}^{A'} = \bar{\iota}^{A'} .$$

Then, for example, given a spinor $Y_{AB'C}$, we can define its dyad components

$$(6.5) \qquad Y_{ab'c} = Y_{AB'C} \zeta_a{}^A \bar{\zeta}_{b'}{}^{B'} \zeta_c{}^C .$$

The lower-case indices behave the same way algebraically as ordinary (capital) spinor indices, but when covariant differentiation is applied, no term involving an affine connection appears for the lower case indices. Thus, the important formal difference between the lower-case and capital indices is simply the difference with respect to covariant differentiation.

Bearing this in mind, it is permissible to choose the components of $\zeta_a{}^A$ to be the Kronecker delta. The dyad components of any spinor will then, in fact, be identical with its spinor components. It is by no means essential to make this specialization but it will be convenient to do so here. The expression

$$(6.6) \qquad \sigma^\mu{}_{ab'} = \sigma^\mu{}_{AB'} \zeta_a{}^A \zeta_{b'}{}^{B'}$$

now gives us l^μ, n^μ, m^μ, $\bar{m}^\mu$ as ab' take, respectively, the values 00', 11', 01', 10', by (6.2). With this interpretation, the expressions

$$(6.7) \qquad \sigma^\mu{}_{00'}, \quad \sigma^\mu{}_{11'}, \quad \sigma^\mu{}_{01'}, \quad \sigma^\mu{}_{10'},$$

may be thought of as giving the tetrad (6.2). However, it is essential to maintain the distinction between the dyad and spinor indices when covariant differentiation is involved.

Explicitly, the covariant derivative $\nabla_\mu \xi_A$ of a spinor ξ_A is

$$(6.8) \qquad \nabla_\mu \xi_A = \partial_\mu \xi_A - \xi_B \Gamma^B{}_{A\mu}$$

where $\Gamma^{B}{}_{A\mu}$ is the spinor affine connection. The corresponding quantity $\overline{\Gamma}^{B'}{}_{A'\mu}$ deals with the primed indices. The rules of covariant differentiation for spinor indices are exactly analogous to those for tensor indices. The choice of $\Gamma^{B}{}_{A\mu}$ is fixed by the requirement that the covariant derivatives of $\sigma^{\mu}{}_{AB'}$, ϵ_{AB}, $\epsilon_{A'B'}$ shall all vanish.[64] (cf. (5.1a).

Since the components of $\zeta_{a}{}^{A}$ are chosen here to be the Kronecker delta:

$$\zeta_{a}{}^{A} = \delta_{a}^{A},$$

it follows that the components of ζ_{aB} are the same as those of ϵ_{AB}. Hence,

$$\nabla_{\mu}\zeta_{aA} = -\zeta_{aC}\Gamma^{C}{}_{A\mu} = \zeta_{a}{}^{C}\Gamma_{CA\mu} = \Gamma_{aA\mu} \tag{6.9}$$

by (6.8). For the analog of the rotation coefficients, we therefore have

$$\Gamma_{abcd'} = \zeta_{b}{}^{A}\,\sigma^{\mu}{}_{cd'}\,\nabla_{\mu}\zeta_{aA} \tag{6.10}$$

with the symmetry

$$\Gamma_{abcd'} = \Gamma_{bacd'}\,. \tag{6.11}$$

These are the <u>spin coefficients</u>

The following notation for the <u>intrinsic derivatives</u> associated with l^{μ}, n^{μ}, m^{μ}, $\overline{m}^{\mu}$ will be used:

$$D \equiv \nabla_{00'}\,,\ \Delta \equiv \nabla_{11'}\,,\ \delta \equiv \nabla_{01'}\,,\ \overline{\delta} \equiv \nabla_{10'}\,. \tag{6.12}$$

The commutation relations for these derivatives acting on scalars:

$$\begin{aligned}\{\nabla_{ab'}\nabla_{cd'} - \nabla_{cd'}\nabla_{ab'}\}\phi \\ \equiv \{\epsilon^{pq}(\Gamma_{pacd'}\nabla_{qb'} - \Gamma_{pcab'}\nabla_{qd'}) \\ + \epsilon^{r's'}(\overline{\Gamma}_{r'b'd'c}\nabla_{as'} - \overline{\Gamma}_{r'd'b'a}\nabla_{cs'})\}\phi\,;\end{aligned} \tag{6.13}$$

CHAPTER 28

can be obtained by direct calculation from (6.6) using (6.10). By a slight extension of this calculation, when the derivatives act on ζ_{aA} we obtain

$$
\begin{aligned}
(6.14)\qquad \nabla_{fe'}\,\Gamma_{acdb'} - \nabla_{db'}\,\Gamma_{acfe'} &= \epsilon^{pq}\{\Gamma_{apdb'}\,\Gamma_{qcfe'} \\
&+ \Gamma_{acpd'}\,\Gamma_{qdfe'} - \Gamma_{apfe'}\,\Gamma_{qcdb'} - \Gamma_{acpe'}\,\Gamma_{qfdb'}\} \\
&+ \epsilon^{r's'}\{\Gamma_{acdr'}\,\overline{\Gamma}_{s'b'e'f} - \Gamma_{acfr'}\,\overline{\Gamma}_{s'e'b'd}\} \\
&+ \Psi_{acdf}\,\epsilon_{e'b'} + \Lambda\,\epsilon_{e'b'}(\epsilon_{cd}\epsilon_{af} + \epsilon_{ad}\epsilon_{cf}) \\
&+ \Phi_{acb'e'}\,\epsilon_{fd}\,,
\end{aligned}
$$

where Ψ_{ABCD}, $\Phi_{ABC'D'}$ and Λ are as in (3.24), (3.25), (3.26), (3.27).

The following twelve Greek letters will be used for the spin coefficients:

(6.15) $\Gamma_{abcd'} =$

cd' \ ab	00	01 or 10	11
00'	κ	ϵ	ϖ
10'	ρ	α	λ
01'	σ	β	μ
11'	τ	γ	ν

In terms of the tetrad vectors, these are $\Gamma_{abcd'} = \frac{1}{2}\epsilon^{p'q'}\sigma^{\mu}_{aq'}\sigma^{\nu}_{cd'}\nabla_{\nu}\sigma_{\mu bp'}$, i.e.

$$
\begin{aligned}
(6.16)\qquad \kappa &= m^{\mu}\ell^{\nu}\nabla_{\nu}\ell_{\mu}\,, & \rho &= m^{\mu}\bar{m}^{\nu}\nabla_{\nu}\ell_{\mu}\,, \\
\sigma &= m^{\mu}m^{\nu}\nabla_{\nu}\ell_{\mu}\,, & \tau &= m^{\mu}n^{\nu}\nabla_{\nu}\ell_{\mu}\,,
\end{aligned}
$$

$$
\begin{aligned}
\epsilon &= \tfrac{1}{2}(n^{\mu}\ell^{\nu}\nabla_{\nu}\ell_{\mu} - \bar{m}^{\mu}\ell^{\nu}\nabla_{\nu}m_{\mu})\,, \\
\alpha &= \tfrac{1}{2}(n^{\mu}\bar{m}^{\nu}\nabla_{\nu}\ell_{\mu} - \bar{m}^{\mu}\bar{m}^{\nu}\nabla_{\nu}m_{\mu})\,,
\end{aligned}
$$

$$\beta = \tfrac{1}{2}\left(n^\mu m^\nu \nabla_\nu \ell_\mu - \bar{m}^\mu m^\nu \nabla_\nu m_\mu \right),$$
$$\gamma = \tfrac{1}{2}\left(n^\mu n^\nu \nabla_\nu \ell_\mu - \bar{m}^\mu n^\nu \nabla_\nu m_\mu \right),$$
$$\varpi = -\bar{m}^\mu \ell^\nu \nabla_\nu n_\mu , \qquad \lambda = -\bar{m}^\mu \bar{m}^\nu \nabla_\nu n_\mu ,$$
$$\mu = -\bar{m}^\lambda m^\nu \nabla_\nu n_\lambda , \qquad \nu = -\bar{m}^\mu n^\lambda \nabla_\lambda n_\mu .$$

The five complex components of Ψ_{ABCD} (i.e. of the Weyl tensor) will be written

$$\Psi_0 = C_{\alpha\beta\gamma\delta}\, \ell^\alpha m^\beta \ell^\gamma m^\delta = \Psi_{0000}$$
$$\Psi_1 = C_{\alpha\beta\gamma\delta}\, \ell^\alpha n^\beta \ell^\gamma m^\delta = \Psi_{0001}$$
(6.17) $$\Psi_2 = C_{\alpha\beta\gamma\delta}\, \ell^\alpha m^\beta \bar{m}^\gamma n^\delta = \Psi_{0011}$$
$$\Psi_3 = C_{\alpha\beta\gamma\delta}\, \ell^\alpha n^\beta \bar{m}^\gamma n^\delta = \Psi_{0111}$$
$$\Psi_4 = C_{\alpha\beta\gamma\delta}\, n^\alpha \bar{m}^\beta n^\gamma \bar{m}^\delta = \Psi_{1111} .$$

Since we shall be concerned only with the Einstein-Maxwell equations here (without cosmological constant) we have (setting $\kappa = 1$ in (4.19))

(6.18) $$\Phi_{abc'd'} = \phi_{ab}\, \bar{\phi}_{c'd'}$$

and we put

$$\phi_0 = F_{\mu\nu} \ell^\mu \ell^\nu = \phi_{00}$$
(6.19) $$\phi_1 = \tfrac{1}{2} F_{\mu\nu} (\ell^\mu n^\nu + \bar{m}^\mu m^\nu) = \phi_{01}$$
$$\phi_2 = F_{\mu\nu} \bar{m}^\mu n^\nu = \phi_{11} .$$

We now make the specialization that the vectors ℓ^μ are the null tangent vectors to a family of (outgoing) null hypersurfaces u = const. In fact, ℓ^μ will be taken to be the gradient of u:

$$\ell^\mu = \nabla^\mu u \quad . \tag{6.20}$$

Furthermore, we choose m^μ and n^μ to be parallelly propagated along the ℓ^μ directions. From (6.16) we then get

$$\kappa = \epsilon = \varpi = 0 \quad , \tag{6.21a}$$

$$\rho = \bar{\rho} \quad , \tag{6.21b}$$

$$\tau = \bar{\alpha} + \beta \quad . \tag{6.21c}$$

We choose coordinates so that $u = x^0$ and so that x^2, x^3 label the geodesics on each hypersurface $u = \text{const}$. Finally we choose $r = x^1$ to be an affine parameter on each of these geodesics. Then

$$\ell_\mu = \delta^0_\mu \, , \quad \ell^\mu = \delta^\mu_1 \tag{6.22}$$

To preserve $\ell_\mu n^\mu = 1$, and $\ell_\mu m^\mu = 0$, we have $(i = 2,3)$

$$m^\mu = \omega \delta^\mu_2 + \xi^i \delta^\mu_i \tag{6.23}$$

$$n^\mu = \delta^\mu_1 + U \delta^\mu_2 + x^i \delta^\mu_i$$

The relation between the tetrad components (6.22) and (6.23) and the metric components is

$$\begin{aligned} g^{22} &= 2 (U - \omega\bar{\omega}), \\ g^{2i} &= x^i - (\xi^i \bar{\omega} + \bar{\xi}^i \omega), \\ g^{ij} &= - (\xi^i \bar{\xi}^j + \bar{\xi}^i \xi^j) , \end{aligned} \tag{6.24}$$

$(i,j = 3, 4)$.

With these simplifications the commutators (6.13) are

$$(\Delta D - D\Delta)\phi = (\gamma+\bar{\gamma})D\phi - \tau\bar{\delta}\phi - \bar{\tau}\delta\phi ,$$

$$(\delta D - D\delta)\phi = \tau D\phi - \sigma\bar{\delta}\phi - \rho\delta\phi ,$$

(6.25)

$$(\delta\Delta - \Delta\delta) = -\bar{\nu} D\phi + \lambda\bar{\delta}\phi + (\mu+\bar{\gamma}-\gamma)\delta\phi ,$$

$$(\bar{\delta}\delta - \delta\bar{\delta}) = (\mu-\bar{\mu})D\phi + (\beta-\bar{\alpha})\bar{\delta}\phi + (\alpha-\bar{\beta})\delta\phi ,$$

with (cf.(6.23), 6.12))

$$D = \partial/\partial r, \quad \delta = \omega\partial/\partial r + \xi^i \partial/\partial x^i ,$$

(6.26)

$$\Delta = U \partial/\partial r + \partial/\partial u + X^i \partial/\partial x^i .$$

In order to relate the tetrad components (or metric components) and the spin coefficients we replace ϕ by u, r, and x^i, respectively, in the four commutators. The result of this operation is (i = 2,3)

(6.27a) $$D\xi^i = \rho\xi^i + \sigma\bar{\xi}^i$$

(6.27b) $$D\omega = \rho\omega + \sigma\bar{\omega} - \tau$$

(6.27c) $$DX^i = \tau\bar{\xi}^i + \bar{\tau}\xi^i$$

(6.27d) $$DU = \tau\bar{\omega} + \bar{\tau}\omega - (\gamma+\bar{\gamma})$$

(6.27e) $$\delta X^i - \Delta\xi^i = (\mu+\bar{\gamma}-\gamma)\xi^i + \bar{\lambda}\bar{\xi}^i$$

(6.27f) $$\delta\bar{\xi}^i - \bar{\delta}\xi^i = (\bar{\beta}-\alpha)\xi^i + (\bar{\alpha}-\beta)\bar{\xi}^i$$

$$\delta\bar{\omega} - \bar{\delta}\omega = (\bar{\beta} - \alpha)\omega + (\bar{\alpha} - \beta)\bar{\omega} + (\mu - \bar{\mu}) \tag{6.27g}$$

$$\delta U - \Delta\omega = (\mu + \bar{\gamma} - \gamma)\omega + \bar{\lambda}\bar{\omega} - \bar{\nu} . \tag{6.27h}$$

The equations (6.14) become

$$D\rho = \rho^2 + \sigma\bar{\sigma} + \phi_0\bar{\phi}_0 \tag{6.28a}$$

$$D\sigma = 2\rho\sigma + \Psi_0 \tag{6.28b}$$

$$D\tau = \tau\rho + \bar{\tau}\sigma + \Psi_1 + \phi_0\bar{\phi}_1 \tag{6.28c}$$

$$D\alpha = \alpha\rho + \beta\bar{\sigma} + \phi_1\bar{\phi}_0 \tag{6.28d}$$

$$D\beta = \beta\rho + \alpha\sigma + \Psi_1 \tag{6.28e}$$

$$D\gamma = \tau\alpha + \bar{\tau}\beta + \Psi_2 + \phi_1\bar{\phi}_1 \tag{6.28f}$$

$$D\lambda = \lambda\rho + \mu\bar{\sigma} + \phi_2\bar{\phi}_0 \tag{6.28g}$$

$$D\mu = \mu\rho + \lambda\sigma + \Psi_2 \tag{6.28h}$$

$$D\nu = \tau\lambda + \bar{\tau}\mu + \Psi_3 + \phi_2\bar{\phi}_1 \tag{6.28i}$$

$$\Delta\lambda - \bar{\delta}\nu = 2\alpha\nu + (\bar{\gamma} - 3\gamma - \mu - \bar{\mu})\lambda - \Psi_4 \tag{6.28j}$$

$$\delta\rho - \bar{\delta}\sigma = \tau\rho + (\bar{\beta} - 3\alpha)\sigma - \Psi_1 + \phi_0\bar{\phi}_1 \tag{6.28k}$$

$$\delta\alpha - \bar{\delta}\beta = \mu\rho - \lambda\sigma - 2\alpha\beta + \alpha\bar{\alpha} + \beta\bar{\beta} - \Psi_2 + \phi_1\bar{\phi}_1 \tag{6.28l}$$

$$\delta\lambda - \bar{\delta}\mu = \bar{\tau}\mu + (\bar{\alpha} - 3\beta)\lambda - \Psi_3 + \phi_2\bar{\phi}_1 \tag{6.28m}$$

$$\delta\nu - \Delta\mu = \gamma\mu - 2\nu\beta + \bar{\gamma}\mu + \mu^2 + \lambda\bar{\lambda} + \phi_2\bar{\phi}_2 \tag{6.28n}$$

$$\delta\gamma - \Delta\beta = \tau\mu - \sigma\nu + (\mu - \gamma + \bar{\gamma})\beta + \bar{\lambda}\alpha + \phi_1\bar{\phi}_2 \tag{6.28o}$$

$$\delta\tau - \Delta\sigma = 2\tau\beta + (\bar{\gamma} + \mu - 3\gamma)\sigma + \bar{\lambda}\rho + \phi_0\bar{\phi}_2 \tag{6.28p}$$

$$\Delta\rho - \bar{\delta}\tau = (\gamma + \bar{\gamma} - \bar{\mu})\rho - 2\alpha\tau - \lambda\sigma - \Psi_2 \tag{6.28q}$$

$$\Delta\alpha - \bar{\delta}\gamma = \rho\nu - \tau\lambda - \lambda\beta + (\bar{\gamma} - \gamma - \bar{\mu})\alpha - \Psi_3 . \tag{6.28r}$$

In the spin coefficient notation with (6.21), Maxwell's equations are

$$(6.29a)\quad D\phi_1 - \bar{\delta}\phi_0 = 2\rho\phi_1 - 2\alpha\phi_0$$

$$(6.29b)\quad D\phi_2 - \bar{\delta}\phi_1 = \rho\phi_2 - \lambda\phi_0$$

$$(6.29c)\quad \delta\phi_1 - \Delta\phi_0 = -\sigma\phi_2 + 2\tau\phi_1 + (\mu - 2\gamma)\phi_0$$

$$(6.29d)\quad \delta\phi_2 - \Delta\phi_1 = (\tau - 2\beta)\phi_2 + 2\mu\phi_1 - \nu\phi_0 ,$$

while the Bianchi identities (3.31) are

$$(6.30a)\quad (D - 4\rho)\Psi_1 - (\bar{\delta} - 4\alpha)\Psi_0 = \bar{\phi}_1 D\phi_0 - \bar{\phi}_0\delta\phi_0 + 2(\bar{\phi}_0\phi_0\beta - \bar{\phi}_0\phi_1\sigma)$$

$$(6.30b)\quad (D - 3\rho)\Psi_2 - (\bar{\delta} - 2\alpha)\Psi_1 + \lambda\Psi_0 = \bar{\phi}_1\bar{\delta}\phi_0 - \bar{\phi}_0\Delta\phi_0 + 2(\bar{\phi}_0\phi_0\gamma - \bar{\phi}_1\phi_0\alpha - \bar{\phi}_0\phi_1\tau + \bar{\phi}_1\phi_1\rho)$$

$$(6.30c)\quad (D - 2\rho)\Psi_3 - \bar{\delta}\Psi_2 + 2\lambda\Psi_1 = \bar{\phi}_1 D\phi_2 - \bar{\phi}_0\delta\phi_2 + 2(\bar{\phi}_0\phi_1\mu - \bar{\phi}_0\phi_2\beta)$$

$$(6.30d)\quad (D - \rho)\Psi_4 - (\bar{\delta} + 2\alpha)\Psi_3 + 3\lambda\Psi_2 = \bar{\phi}_1\bar{\delta}\phi_2 - \bar{\phi}_0\Delta\phi_2 + 2(\bar{\phi}_0\phi_1\nu - \bar{\phi}_1\phi_1\lambda - \bar{\phi}_0\phi_2\gamma + \bar{\phi}_1\phi_2\alpha)$$

$$(6.30e)\quad (\delta - 4\tau - 2\beta)\Psi_1 + 3\sigma\Psi_2 - (\Delta - 4\gamma + \mu)\Psi_0 = \bar{\phi}_2 D\phi_0 - \bar{\phi}_1\delta\phi_0 + 2(\bar{\phi}_1\phi_0\beta - \bar{\phi}_1\phi_1\sigma)$$

$$(6.30f)\quad (\delta - 3\tau)\Psi_2 + 2\sigma\Psi_3 - (\Delta - 2\gamma + 2\mu)\Psi_1 + \nu\Psi_0 = \bar{\phi}_2\bar{\delta}\phi_0 - \bar{\phi}_1\Delta\phi_0 + 2(\bar{\phi}_1\phi_0\gamma - \bar{\phi}_2\phi_0\alpha - \bar{\phi}_1\phi_1\tau + \bar{\phi}_2\phi_1\rho)$$

$$(6.30g)\quad (\delta - 2\tau + 2\beta)\Psi_3 + \sigma\Psi_4 - (\Delta + 3\mu)\Psi_2 + 2\nu\Psi_1$$
$$= \bar{\phi}_2 \,\mathrm{D}\phi_2 - \bar{\phi}_1 \delta\phi_2 + 2(\bar{\phi}_1\phi_1\mu - \bar{\phi}_1\phi_2\beta)$$

$$(6.30h)\quad (\delta - \tau + 4\beta)\Psi_4 - (\Delta + 2\gamma + 4\mu)\Psi_3 + 3\nu\Psi_2$$
$$= \bar{\phi}_2\bar{\delta}\phi_2 - \bar{\phi}_1\Delta\phi_2 + 2(\bar{\phi}_1\phi_1\nu - \bar{\phi}_2\phi_1\lambda - \bar{\phi}_1\phi_2\gamma + \bar{\phi}_2\phi_2\alpha).$$

The quantities Ψ_0 and ϕ_0 are the respective gravitational and electromagnetic null-data for a hypersurface u = const. We may therefore expect that the nature of the solution will be governed by the behaviour of these two quantities. We require to impose a condition of approach to flatness at infinity along u = const. From a study of the linear theories and, in particular, those solutions involving outgoing waves, the conditions

$$(6.31)\qquad \Psi_0 = 0\,(r^{-5}),\quad \phi_0 = 0\,(r^{-3})$$

emerge as reasonable ones to impose. In fact, from the conditions (6.31), together with a suitable number of r, x^2, x^3 derivatives of (6.31) it is possible to obtain the asymptotic behaviour of all the remaining components of Ψ_{ABCD} and ϕ_{AB} and of all the spin coefficients [59, 60] The somewhat stronger versions

$$(6.32a)\qquad \Psi_0 = \psi_0^0\, r^{-5} + \psi_0^1\, r^{-6} + 0\,(r^{-7})$$

$$(6.32b)\qquad \phi_0 = \phi_0^0\, r^{-3} + \phi_0^1\, r^{-4} + 0\,(r^{-5})\ ,$$

where the ψ_0^n and ϕ_0^n are independent of r, together with the assumptions that a suitable number of r, u, x^2, x^3 derivatives of these expressions also hold, will be required here.

By suitably choosing the coordinate system a number of simplifications in the asymptotic behaviour of various spin coefficients can be obtained. Also, we may choose x^2 and x^3 to be asymptotically spherical polar coordinates for

the "sphere at infinity" on u = const. Finally, the operator $\eth$ can be introduced to obtain a great improvement in the appearance of the expressions. It will be useful to digress slightly, therefore, so as to describe some of the main properties of the $\eth$ operator.

Consider a rotation of the spacelike vectors Re (m^μ), Im (m^μ) in their plane given by

$$(6.33) \qquad (m^\mu)' = e^{i\chi} m^\mu$$

(with $(l^\mu)' = l^\mu$, χ real). A quantity η will be said to have spin weight s if it transforms as

$$(6.34) \qquad \eta' = e^{si\chi} \eta$$

under (6.33) (cf. (5.15)).

If spherical polar coordinates are used, a natural choice [65,63] for m^μ is to make Re (m^μ) and Im (m^μ) tangential, respectively, to the curves φ = const and θ = const. Another convenient coordinate system for the sphere is (ζ, $\bar{\zeta}$) where the complex parameter ζ is related to (θ, φ) by

$$(6.35) \qquad \zeta = e^{i\varphi} \cot \tfrac{1}{2}\theta \ .$$

In this case, the natural choice for m^μ is to make Re (m^μ) and Im (m^μ) tangential, respectively, to the curves Im (ζ) = const and Re (ζ) = const.

Let η be a quantity defined on the sphere of spin weight s. Define the operator $\eth$, in a particular (θ, φ) coordinate system, by [63]

$$(6.36) \qquad \eth \eta = - (\sin\theta)^s \left\{ \frac{\partial}{\partial\theta} + \frac{i}{\sin\theta} \frac{\partial}{\partial\varphi} \right\} \{ (\sin\theta)^{-s} \eta \} \ .$$

(The operator $\eth$ is effectively a covariant differentiation operator in the surface.) Under a rotation of m^μ (but, for the moment, keeping the coordinates fixed) we demand that $\eth\eta$ behave as a quantity of spin weight s + 1. From

(6.35) and the fact that the m^μ vectors are rotated through an angle $-\varphi$ in the passage from (θ, φ) to $(\zeta, \bar{\zeta})$, we obtain

$$\text{(6.37)} \qquad \eth \eta = P^{1-s} \partial (P^s \eta) / \partial \zeta$$

as the definition of $\eth$ in the $(\zeta, \bar{\zeta})$ system, where

$$\text{(6.38)} \qquad P = (1 + \zeta\bar{\zeta}).$$

Now, the fundamental property of the operator $\eth$ is that it is invariant (with spin weight unity) under change of coordinate system which preserves the sphere metric

$$\text{(6.39)} \qquad ds^2 = d\theta^2 + \sin^2\theta \; d\varphi^2 = 4P^{-2} \, d\zeta \, d\bar{\zeta} .$$

This is most easily seen in the $(\zeta, \bar{\zeta})$ system and the result for the (θ, φ) system will then follow. With χ as in (6.33), so that $\eta' = e^{si\chi} \eta$, we have

$$\text{(6.40)} \qquad e^{-2i\chi} = \frac{\partial \zeta' / \partial \zeta}{\partial \bar{\zeta}' / \partial \bar{\zeta}} ,$$

where $(\zeta', \bar{\zeta}')$ is the new coordinate system, with new complex tangent vector m'^μ Since $ds' = ds$, we have

$$\text{(6.41)} \qquad \partial \zeta' / \partial \zeta = e^{-i\chi} \, P'/P$$

with $\partial \zeta' / \partial \bar{\zeta} = 0$. From this it follows that $\eth' \eta' = P'^{1-s} \partial (P'^s \eta') / \partial \zeta' = e^{(s+1)i\chi} \eth \eta$ as required. The operator $\bar{\eth}$ is defined in the $(\zeta, \bar{\zeta})$ system by

$$\text{(6.42)} \qquad \bar{\eth} \eta = P^{1+s} \partial (P^{-s} \eta) / \partial \bar{\zeta} .$$

The operators $\eth$ and $\bar{\eth}$ do not in general commute. In fact, we have

$$\text{(6.43)} \qquad (\bar{\eth} \eth - \eth \bar{\eth}) \eta = 2s\eta .$$

(However, it is a consequence of the following considerations that $\eth^p \bar{\eth}^q \eta = \bar{\eth}^q \eth^p \eta$ whenever $q - p = 2s$.)

Another important property of $\eth$ concerns its effect on spherical harmonics. Let ${}_0Y_{\ell,m}$ ($\ell = 0, 1, \ldots ; m = -\ell, \ldots, +\ell$) be the usual "scalar" spherical harmonics. Then we can define the spin s spherical harmonics (for integral spin) as follows:

$$(6.44)\qquad {}_sY_{\ell,m} = \begin{cases} \left[\frac{(\ell - s)!}{(\ell + s)!}\right]^{\frac{1}{2}} \eth^s {}_0Y_{\ell,m} & (0 \leqslant s \leqslant \ell), \\ (-1)^s \left[\frac{(\ell + s)!}{(\ell - s)!}\right]^{\frac{1}{2}} \bar{\eth}^{-s} {}_0Y_{\ell,m} & (-\ell \leqslant s \leqslant 0). \end{cases}$$

The ${}_sY_{\ell,m}$ are not defined for $|s| > \ell$. In the $(\zeta, \bar{\zeta})$ system, the spin s harmonics take the form (which also holds when s, ℓ, m are all half-odd integers):

$$(6.45)\qquad {}_sY_{\ell,m} = \frac{(-1)^{\ell - m}[(\ell + m)!(\ell - m)!(2\ell + 1)]^{\frac{1}{2}}}{[4\pi(\ell - s)!(\ell + s)!]^{1/2}} (1 + \zeta\bar{\zeta})^{-\ell}$$

$$\times \sum_p \zeta^p (-\bar{\zeta})^{p+s-m} \binom{\ell - s}{p} \binom{\ell + s}{p + s - m},$$

summed over integral values of p. To express these in the (θ, φ)-system, we multiply by $e^{is\varphi}$ and replace ζ according to (6.35). By (6.45) we have ${}_s\bar{Y}_{\ell,m} = (-1)^{m+s} {}_{-s}Y_{\ell,-m}$ and, for all s with $|s| \leqslant \ell$,

$$(6.46)\qquad \eth({}_sY_{\ell,m}) = [(\ell - s)(\ell + s + 1)]^{1/2} {}_{s+1}Y_{\ell,-m}$$

$$\bar{\eth}({}_sY_{\ell,m}) = -[(\ell + s)(\ell - s + 1)]^{1/2} {}_{s-1}Y_{\ell,m}.$$

In particular, $\eth$ annihilates ${}_sY_{\ell,m}$ whenever $\ell = s$ and $\bar{\eth}$ annihilates ${}_sY_{\ell,m}$ if $\ell = -s$. We see, furthermore, that $\eth^p$ annihilates any quantity of spin weight s which is composed only of harmonics with $\ell < s + p$.

The ${}_sY_{\ell,m}$ are eigenfunctions of the operator $\bar{\eth}\eth$ for each spin weight s:

$$(6.47)\qquad \bar{\eth}\eth({}_sY_{\ell,m}) = -(\ell - s)(\ell + s + 1) {}_sY_{\ell,m}.$$

(If $s = 0$, $\bar{\eth}\eth$ is essentially the total angular momentum operator.)

For each given s, the ${}_sY_{\ell,m}$ form a set of orthonormal functions of spin weight s on the sphere:

$$\int {}_sY_{\ell,m}\ \bar{{}_sY}_{\ell',m'}\, dS = \delta_{\ell\ell'}\ \delta_{mm'}\ , \tag{6.48}$$

dS being the surface area element on the unit sphere. This can be proved (if s is an integer) by induction on s using (6.46), the orthonormality of $Y_{\ell,m}$ and the fact that

$$\int (\eth A)B\, dS = -\int A(\eth B)\, dS \tag{6.49}$$

with (spin weight of A) + (spin weight of B) = -1. It is immediate from (6.36) that $\int \eth C\, dS = 0$ (where C has spin weight -1) and (6.49) follows. The orthonormal functions ${}_sY_{\ell,m}$ are also complete, for spin weight s quantities on the sphere. That is, if η has spin weight s and is suitably regular on the sphere, there exist constants $\eta_{\ell,m}$ such that

$$\eta = \sum_{\ell,m} \eta_{\ell,m}\ {}_sY_{\ell,m}\ . \tag{6.50}$$

Another way of expressing this result is: given any suitably regular η on the sphere, of (integral) spin weight $s > 0$, there exists ξ of spin weight zero for which

$$\eta = \eth^s \xi\ . \tag{6.51}$$

We are now in a position to write down the equations for the asymptotic fields and spin coefficients which are the consequences of (6.32). In fact, more terms could be calculated, if desired, by substituting into the equations (6.28), (6.29), (6.30). We have

$$\Psi_1 = \Psi_1^0\, r^{-4} + (\bar{\eth}\, \Psi_0^0 + 3\, \bar{\phi}_1^0 \phi_0^0)\, r^{-5} + O(r^{-6}) \tag{6.52a}$$

$$\Psi_2 = \Psi_2^0\, r^{-3} + (\bar{\eth}\, \Psi_1^0 + 2\, \bar{\phi}_1^0 \phi_1^0)\, r^{-4} + O(r^{-5}) \tag{6.52b}$$

(6.52c) $$\psi_3 = \psi_3^0\, r^{-2} + (\bar{\eth}\psi_2^0 + \bar{\phi}_1^0\phi_2^0)\, r^{-3} + 0(r^{-4})$$

(6.52d) $$\psi_4 = \psi_4^0\, r^{-1} + \bar{\eth}\psi_3^0\, r^{-2} + 0(r^{-3})$$

(6.52e) $$\phi_1 = \phi_1^0 r^{-2} + \bar{\eth}\phi_0^0\, r^{-3} + 0(r^{-4})$$

(6.52f) $$\phi_2 = \phi_2^0\, r^{-1} + \bar{\eth}\phi_1^0\, r^{-2} + 0(r^{-3})$$

(6.53a) $$\rho = -r^{-1} - \sigma^0\bar{\sigma}^0\, r^{-3} + 0(r^{-5})$$

(6.53b) $$\sigma = \sigma^0\, r^{-2} + (\bar{\sigma}^0\sigma^{0^2} - \tfrac{1}{2}\psi_0^0)\, r^{-4} + 0(r^{-5})$$

(6.53c) $$\alpha = \alpha^0\, r^{-1} + \bar{\sigma}^0\bar{\alpha}^0\, r^{-2} + \sigma^0\bar{\sigma}^0\,\alpha^0\, r^{-3} + 0(r^{-4})$$

(6.53d) $$\beta = -\bar{\alpha}^0\, r^{-1} - \sigma^0\alpha^0\, r^{-2} - (\sigma^0\bar{\sigma}^0\bar{\alpha}^0 + \tfrac{1}{2}\psi_1^0) r^{-3} + 0(r^{-4})$$

(6.53e) $$\tau = -\tfrac{1}{2}\psi_1^0\, r^{-3} + \tfrac{1}{3}(\tfrac{1}{2}\sigma^0\bar{\psi}_1^0 - \bar{\eth}\psi_0^0 - 4\phi_0^0\bar{\phi}_1^0)\, r^{-4} + 0(r^{-5})$$

(6.53f) $$\lambda = \dot{\bar{\sigma}}^0\, r^{-1} + \bar{\sigma}^0\, r^{-2} + (\sigma^0\bar{\sigma}^0\dot{\bar{\sigma}}^0 + \tfrac{1}{2}\bar{\sigma}^0\psi_2^0 - \tfrac{1}{2}\bar{\phi}_0^0\phi_2^0)\, r^{-3} + 0\,(r^{-4})$$

(6.53g) $$\mu = -\,r^{-1} - (\sigma^0\dot{\bar{\sigma}}^0 + \psi_2^0) r^{-2} - (\sigma^0\bar{\sigma}^0 + \tfrac{1}{2}\bar{\eth}\psi_1^0 + \phi_1^0\bar{\phi}_1^0)\, r^{-3} + 0\,(r^{-4})$$

(6.53h) $$\gamma = -\tfrac{1}{2}\psi_2^0\, r^{-2} - \tfrac{1}{6}(2\bar{\eth}\psi_1^0 - \alpha^0\psi_1^0 + \bar{\alpha}^0\bar{\psi}_1^0 + 6\phi_1^0\bar{\phi}_1^0)\, r^{-3} + 0(r^{-4})$$

(6.53i) $$\nu = -\psi_3^0\, r^{-1} - (\tfrac{1}{2}\bar{\eth}\psi_2^0 + \bar{\phi}_1^0\phi_2^0)\, r^{-2} + 0(r^{-3})$$

(6.54a) $$U = -1 - \tfrac{1}{2}(\psi_2^0 + \bar{\psi}_2^0) r^{-1} - \tfrac{1}{6}(\bar{\eth}\psi_1^0 + \eth\bar{\psi}_1^0 + 6\phi_1^0\bar{\phi}_1^0) r^{-2} + 0(r^{-3})$$

6.54b) $$X^3 + i\,X^4 \sin\theta = \tfrac{1}{3}\psi_1^0\, r^{-3} + 0\,(r^{-4})$$

X^3 and X^4 being real,

(6.54c) $$\xi^3 = r^{-1} - \sigma^0\, r^{-2} + \sigma^0\bar{\sigma}^0\, r^{-3} + 0(r^{-4})$$

(6.54d) $$\xi^4 = \frac{i}{\sin\theta}\, r^{-1} + \frac{i\,\sigma^0}{\sin\theta}\, r^{-2} + \frac{i\,\sigma^0\sigma^0}{\sin\theta}\, r^{-3} + 0(r^{-4})$$

$$\omega = -\bar{\eth}\,\sigma^0\, r^{-1} + (\sigma^0\eth\bar{\sigma}^0 - \tfrac{1}{2}\psi_1^0)\, r^{-2} + 0\,(r^{-3}) \tag{6.54e}$$

where

$$\alpha^0 = -\tfrac{1}{2}\cot\theta \tag{6.55}$$

Of these equations, only (6.54b), (6.54c), (6.54d) and (6.55) depend on the explicit use of spherical polar coordinates and corresponding m^μ vectors. We have the further equations involving u-derivatives, with the dot denoting $\partial/\partial u$:

$$\psi_2^0 - \bar{\psi}_2^0 = \bar{\eth}^2\sigma^0 - \eth^2\,\bar{\sigma}^0 + \bar{\sigma}^0\dot{\sigma}^0 - \sigma^0\dot{\bar{\sigma}}^0 \tag{6.56a}$$

$$\psi_3^0 = \eth\,\dot{\bar{\sigma}}^0 \tag{6.56b}$$

$$\psi_4^0 = -\ddot{\bar{\sigma}}^0 \tag{6.56c}$$

$$\dot{\psi}_0^0 + \eth\,\psi_1^0 - 3\,\sigma^0\psi_2^0 - 3\phi_0^0\bar{\phi}_2^0 = 0 \tag{6.57a}$$

$$\dot{\psi}_0^1 - 4\,\eth(\sigma^0\psi_1^0) + \bar{\eth}\eth\psi_0^0 + 4\,\bar{\phi}_1^0\,\eth\,\phi_0^0 - 8\,\sigma^0\phi_1^0\bar{\phi}_1^0 - 4\phi_0^1\bar{\phi}_2^0 = 0 \tag{6.57b}$$

$$\dot{\psi}_1^0 + \eth\psi_2^0 - 2\,\sigma^0\psi_3^0 - 2\phi_1^0\bar{\phi}_2^0 = 0 \tag{6.57c}$$

$$\dot{\psi}_2^0 + \eth\psi_3^0 - \sigma^0\psi_4^0 - \phi_2^0\bar{\phi}_2^0 = 0 \tag{6.57d}$$

$$\dot{\phi}_0^0 + \eth\phi_1^0 - \sigma^0\phi_2^0 = 0 \tag{6.57e}$$

$$\dot{\phi}_0^1 + \bar{\eth}\eth\phi_0^0 - 2\,\bar{\eth}(\sigma^0\phi_1^0) = 0 \tag{6.57f}$$

$$\dot{\phi}_1^0 + \eth\,\phi_2^0 = 0 \tag{6.57g}$$

There is a particular result concerning $\eth$ which is very useful in proving conservation laws. Let η be any quantity of spin-weight $-1-\ell$, which is suitably regular on the (θ, φ)-sphere. Then, for $\ell \geqslant |s|$,

$$(6.58) \qquad \int {}_sY_{\ell,m} \; \eth^{\ell-s+1} \eta \; dS = 0.$$

($dS = d\theta \sin\theta \, d\varphi$.) To prove (6.58), we need only to integrate repeatedly, by parts, according to (6.47) and use (6.46): $\eth^{\ell-s+1} \, {}_sY_{\ell,m} = 0$. A very simple application of (6.58) is the law of conservation of total charge. For the quantity

$$(6.59) \qquad e = \int {}_0Y_{0,0} \, \phi_1^0 \; dS$$

is proportional to the total charge (+ i × "magnetic charge"). (Of course, ${}_0Y_{0,0} = \frac{1}{2}\pi^{-\frac{1}{2}}$.) Now, by (6.57g), $\dot{e} = \int {}_0Y_{0,0} \, \dot{\phi}_1^0 \, dS = -\int {}_0Y_{0,0} \, \eth \phi_2^0 \, dS = 0$, by (6.58).

In a similar way, we can obtain the Bondi-Sachs mass-momentum conservation law, [61] for which we put

$$(6.60) \qquad P_{n,m} = -(32\pi)^{-1/2} \int {}_0Y_{n,m} \, (\sigma^0 \dot{\bar{\sigma}}^0 + \Psi_2^0) \; dS.$$

(The quantity $\dot{\sigma}^0$ is essentially the Bondi-Sachs complex news function, σ^0 being essentially the Bondi-Sachs "c".) The quantity whose components in an "asymptotic orthonormal frame" are

$$(6.61) \qquad \left(P_{0,0}, \; 6^{-1/2}(P_{1,1} - P_{1,-1}), \; (-6)^{-1/2}(P_{1,1} + P_{1,-1}), \; 3^{-1/2} P_{1,0}\right)$$

is the Bondi-Sachs total energy-momentum 4-vector at the retarded time u, with $P_{0,0}$ being the total mass component. This vector is in fact real even though $\sigma^0 \dot{\bar{\sigma}}^0 + \Psi_2^0$ is complex. This vanishing of the imaginary parts obtained from the integrals, follows by (6.58) because of (6.56a). Now, to obtain the conservation law, we simply use (6.57d):

$$(6.62)\qquad \dot{P}_{n,m} = -(32\pi)^{-1/2}\int_{0}Y_{n,m}\left(|\dot{\sigma}^{0}|^{2} - \eth^{2}\,\dot{\bar{\sigma}}^{0} + |\phi_{2}^{0}|^{2}\right) dS$$

$$= -Q_{n,m} - R_{n,m}$$

by (6.58), where

$$(6.63a)\qquad Q_{n,m} = (32\pi)^{-1/2}\int_{0}Y_{n,m}\,|\dot{\sigma}^{0}|^{2}\,dS$$

and

$$(6.63b)\qquad R_{n,m} = (32\pi)^{-1/2}\int_{0}Y_{n,m}\,|\phi_{2}^{0}|^{2}\,dS.$$

The asymptotic 4-vectors corresponding to (6.63a) and (6.63b) as in (6.61), are both timelike and future-pointing (or zero). The energy-momentum carried away by gravitational radiation and by electromagnetic radiation, respectively, between the retarded times u and $u + du$ are given by the asymptotic 4-vectors corresponding to $Q_{n,m}\,du$ and $R_{n,m}\,du$. The Bondi mass-loss theorem is a special case of this.

Finally, let us consider the new "absolute" conservation laws.[62] First, for the case of the electromagnetic conserved quantities we put

$$(6.64)\qquad F_{m} = \int_{1}\bar{Y}_{1,m}\,\phi_{0}^{1}\,dS.$$

These are three complex quantities ($m = -1,0,1$) and therefore equivalent to six real ones. The quantities are conserved in the sense $\dot{F}_{m} = 0$, as follows at once from (6.57e) and (6.58), since $\dot{\phi}_{0}^{1} = \eth(2\sigma^{0}\phi_{1}^{0} - \eth\phi_{0}^{0})$. When the electromagnetic field vanishes, we have $\dot{\psi}_{0}^{1} = \bar{\eth}(4\sigma^{0}\psi_{1}^{0} - \eth\psi_{0}^{0})$, by (6.57b), so the same reasoning tells us that the five complex (or ten real) quantities

$$(6.65)\qquad G_{m} = \int {}_{2}\bar{Y}_{2,m}\,\psi_{0}^{1}\,dS$$

are also conserved in the sense $G_m = 0$. When the electromagnetic field does not vanish, we can preserve this conservation law by adding a correction term to (6.65). This is given by

$$(6.66)\qquad G_m = \int {}_2\bar{Y}_{2,m} \left[\psi_0^1 + 4F \bar{\eth}^{-1} \left(\bar{\phi}_1^0 - E \right) - 4 \bar{\phi}_1^0 \bar{\eth}^{-1} \left(\phi_0^1 - F \right) \right] dS$$

(or other equivalent expressions), where

$$(6.67)\qquad E = \bar{e}\ {}_0Y_{0,0},\quad F = \sum_{m=-1}^{1} F_m\ {}_1Y_{1,m}.$$

Exton has found an additional correction term which preserves $\dot{G}_m = 0$ even in the presence of neutrino fields.

An identification of the G_m and the F_m in terms of multipole moments can also be achieved using this formalism, but the (somewhat complicated) expressions for $\dot{\psi}_0^2$ and $\dot{\phi}_0^2$ analogous to those of (6.57) are required.

VII. THE CONFORMAL TECHNIQUE

The formalism of section VI is very convenient for performing many detailed calculations, but it is not so well-suited for the discussion of general invariance properties. For example, while conservation laws were readily obtained, the conservation was always expressed in terms of the dependence of quantities on a particular retarded time parameter u. There is a more general type of conservation law which applies to asymptotically flat space-times. For, we may be interested in comparing a quantity defined with respect to two different outgoing null hypersurfaces, which may not both be describable as u = const. hypersurfaces of the same allowable coordinate system. What is required, in fact is the examination of the behaviour of quantities under the full Bondi-Metzner-Sachs group.[61,67]

To discuss such matters it is convenient to use a totally different formalism which is much more evidently geometrical in character.[46,66] It enables, in particular, a completely geometrical definition to be given, which is effectively equivalent to the asymptotic flatness requirements of section VI. (These, in turn, are essentially equivalent to the asymptotic conditions of the Bondi-Sachs type of formalism, which differs mainly in that a somewhat different type of coordinate system is used.) The geometrical approach that I shall describe, can in fact be used to obtain the complete behaviour of such quantities as the $P_{n,m}$, the F_m and the G_m of section VI. (I shall not give the arguments here, however.) The approach has the additional advantage in that it is not biased with respect to a particular choice of either outgoing or incoming null hypersurfaces. Thus, incoming and outgoing radiation fields can be considered simultaneously and a precise definition of these radiation fields can be given. This would appear to give a useful framework for the discussion of scattering problems, etc. The approach can, in fact, be used to discuss asymptotic questions of a more general character than those which just apply to asymptotically flat space-times. Some of these matters will now be discussed.

The study of asymptotic behaviour is the study of behaviour in the 'neighbourhood of infinity'. Actual points <u>at</u> 'infinity' are, however, not generally thought of as having any consistently sensible existence. The main idea here is to formalize a concept of 'points at infinity' so that they can, in fact, be introduced in a rigorous and consistent way. These points at 'infinity' will form a kind of hypersurface boundary to space-time. 'Asymptotic behaviour' (as considered here) will then turn out to be essentially no different from behaviour at any finite hypersurface. The basic tool is provided by the concept of a <u>conformal transformation</u> of space-time. Such transformations preserve <u>ratios</u> of local infinitesimal distances, or equivalently, the positions of the light cones. As a consequence, it turns out that the essential features of zero rest-mass fields are preserved, and their asymptotic behaviour may be studied by using <u>local</u> differential techniques in the way indicated above.

One starting-point in the study of the conformal structure of infinity for <u>Minkowski</u> space-time is provided by the <u>inversion</u>:

$$(7.1)\qquad x^\mu = -\,\hat{x}^\mu(\hat{x}_\nu\,\hat{x}^\nu)^{-1},\quad \hat{x}^\mu = -\,x^\mu(x_\nu\,x^\nu)^{-1};$$

$$g_{\mu\nu} = \hat{g}_{\mu\nu} = \mathrm{diag}\,\{+1, -1, -1, -1\}$$

This is conformal since

$$\hat{g}_{\mu\nu}\,d\hat{x}^\mu\,d\hat{x}^\nu = \Omega^2 g_{\mu\nu}\,dx^\mu dx^\nu,$$

where

$$\Omega = \hat{x}_\mu\,\hat{x}^\mu = (x_\mu\,x^\mu)^{-1}.$$

The null cone of the $\hat{x}^\mu$ origin is mapped to x^μ infinity and $\hat{x}^\mu$ infinity is mapped into the null cone of the x^μ origin. (The minus sign in (7.1) is to preserve time sense.) We may regard (7.1) either as giving a conformal

point transformation in Minkowski space-time which maps infinity into a finite null cone, or else, as a relation between two co-ordinate systems, each of which covers part of a certain highly symmetrical (compact) conformal manifold $\mathscr{R}$, the 'compactification' of Minkowski space-time[68]. These two co-ordinate systems do not quite cover the whole of $\mathscr{R}$ there being a additional sphere of points lying at infinity on the null cones of both origins. But these remaining points are easily covered by choosing a third co-ordinate system which is related to $x^\mu - a^\mu$ in the same way as $\hat{x}^\mu$ is related to x^μ, where a^μ is a fixed time-like vector. The resulting manifold $\mathscr{R}$ has many interesting properties, but they will not be discussed here, the essential relevant feature being that $\mathscr{R}$ is obtained from Minkowski space-time by the addition of a (closed) null cone $\mathscr{I}$ at infinity. From the point of view of conformal structure, $\mathscr{I}$ is just one null cone on an equal footing with any other null cone in $\mathscr{R}$ (all of which are closed).

There is one property of $\mathscr{R}$, however, which seems undesirable when these ideas are applied to interacting fields, or to curved space-times. This is the fact that the 'future infinity' turns out to have been identified with the past infinity' in the definition of $\mathscr{R}$. To avoid this feature it will be desirable effectively to 'cut' this manifold along the hypersurface $\mathscr{I}$ and to consider instead the resulting manifold with boundary. This boundary consists essentially of two copies of $\mathscr{I}$, one in the 'future' which will be called $\mathscr{I}^+$ and one in the 'past' to be called $\mathscr{I}^-$ (see figure 2). The four-dimensional region between these two bounding hypersurfaces corresponds to the original Minkowski space-time. The hypersurface $\mathscr{I}^-$ is a 'closed up' null cone with a past vertex I^- and also a future vertex I^0. This point I^0 is also the past vertex of the null cone $\mathscr{I}^+$ and $\mathscr{I}^+$ also has a future vertex I^+. For preciseness, I shall exclude the three points I^-, I^0, I^+ from the definitions of the sets $\mathscr{I}^-$, $\mathscr{I}^+$.

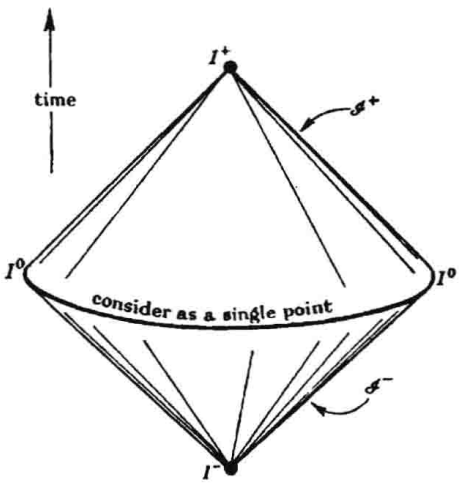


FIGURE 2. Minkowski space-time can be mapped onto the interior of a region bounded by two cones joined base to base. The 'equator' at which the cones are joined should be considered as a single point. To obtain the manifold $\mathscr{R}$, opposite generators must be identified with future sense preserved.

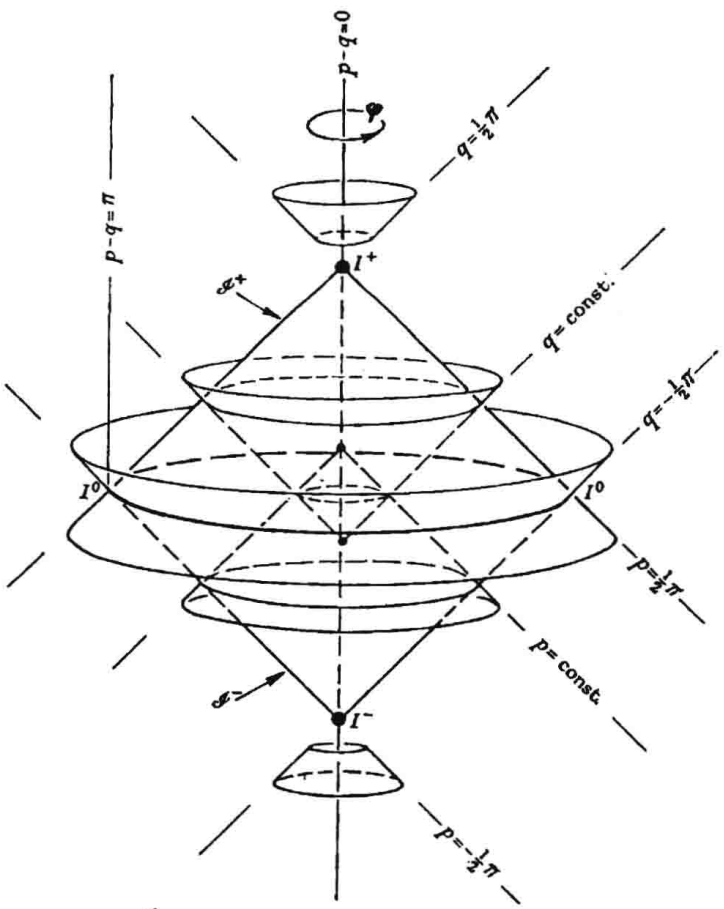


Fig. 3. Coordinates for $\hat{\mathscr{M}}$ in Minkowski case. The θ coordinate has been suppressed ($\theta = \pi/2$).

In order to get a clearer picture of the situation let us consider null coordinates u, v (advanced and retarded times) in Minkowski space-time $\mathcal{M}$:

$$(7.2) \qquad ds^2 = dudv - \tfrac{1}{4}(u - v)^2 \{ d\theta^2 + \sin^2\theta \, d\varphi^2 \}$$

$u \geqslant v$. A convenient conformal factor is given by

$$(7.3) \qquad \Omega = (1 + u^2)^{-1/2} (1 + v^2)^{-1/2} .$$

Changing coordinates by

$$u = \tan p, \quad v = \tan q,$$

the new "unphysical" metric, for the "compactified" manifold $\hat{\mathcal{M}}$, takes the form:

$$(7.4) \qquad d\hat{s}^2 = \Omega^2 ds^2 = dpdq - \tfrac{1}{4}\sin^2(p - q) \{ d\theta^2 + \sin^2\theta \, d\varphi^2 \}$$

where p and q vary from $-\pi/2$ to $\pi/2$ with $p \geqslant q$. The hypersurface boundary $\mathscr{I}$ of $\hat{\mathcal{M}}$ is given when $q = -\pi/2$ or $p = \pi/2$, since these values correspond to infinity ($v = -\infty$ or $u = \infty$) for $\mathcal{M}$. Denote the region given by $q = -\pi/2$, $-\pi/2 < p < \pi/2$ by $\mathscr{I}^-$. Similarly let $\mathscr{I}^+$ be given by $p = \pi/2$, $-\pi/2 < q < \pi/2$. The point given by $p = q = -\pi/2$ is denoted by I^-, $p = q = \pi/2$ by I^+ and $p = -q = \pi/2$ by I° (see Fig. 3). The fact that I^-, I^+ and I° are <u>single points</u> follows from the fact that $\sin(p - q) = 0$ there. These are coordinate singularities of the type encountered at the origin of polar coordinates. The manifold $\hat{\mathcal{M}}$ is regular at these points. In fact, the metric $d\hat{s}$ is that of an Einstein static universe -- a cylinder ($S^3 \times E^1$) which represents a three-dimensional spherical space which is constant in time. The manifold $\hat{\mathcal{M}}$ is only a finite portion of this cylinder (see Fig. 4). The hypersurface $\mathscr{I}^-$ is a future null cone with vertex I^-. This cone is refocused to a point I° which is spacially diametrically opposite to I^-. The future null cone of I° is $\mathscr{I}^+$ which is focused again at I^+. The region lying between $\mathscr{I}^-$ and $\mathscr{I}^+$ is conformal to the entire Minkowski space-time.

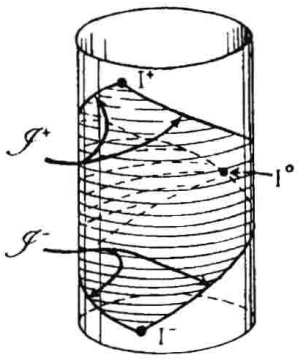


Fig. 4. The cylinder represents the Einstein universe. The two spatial coordinates θ, φ have been suppressed; only one space and one time dimension are depicted. The shaded region corresponds to the manifold $\mathscr{M}$. (Minkowski case.)

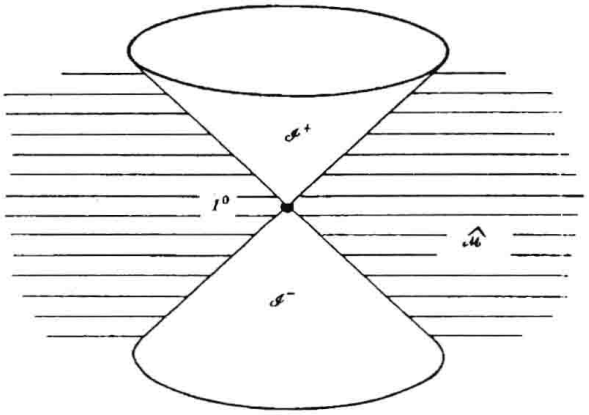


FIGURE 5. The neighbourhood of I^0. Physical space-time is represented by the shaded region. (Minkowski case.)

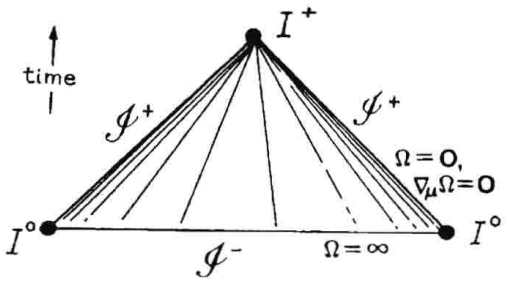


Fig. 6. $\mathscr{M}$ for the Einstein–de Sitter model ($\lambda = 0$, $k = 0$), with the world-lines of the galaxies drawn in. I^0 is to be a single point. The initial singularity is represented by $\mathscr{I}^-$. The relativistic cosmology with $\lambda = 0$, $k < 0$ is similar except that I^0 is an S^2. The corresponding cases with $\lambda \neq 0$ look similar, except that $\mathscr{I}^+$ is spacelike and I^+ is absent.

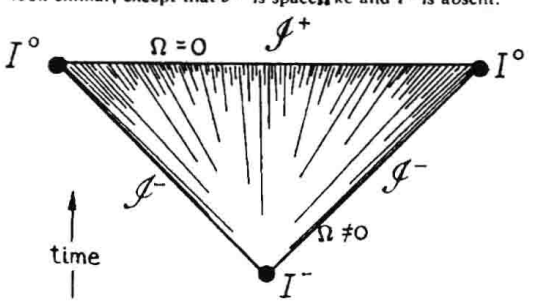


Fig. 7. $\mathscr{M}$ for the steady-state model showing the world-lines of the galaxies. Again I^0 is a single point.

The meaning of I^-, $\mathscr{I}^-$, I°, $\mathscr{I}^+$, I^+, can be seen by considering the behaviour of curves in $\hat{\mathscr{M}}$ corresponding to the straight lines in $\mathscr{M}$. The image of a time-like straight line originates at I^- and terminates at I^+; the image of a space-like straight line originates and terminates at I°; the image of a null straight line originates at a point of $\mathscr{I}^-$ and terminates at a point of $\mathscr{I}^+$. Thus, at I^- represents past infinity; I° represents spatial infinity; I^+ represents future infinity; $\mathscr{I}^-$ represents <u>past null</u> infinity; $\mathscr{I}^+$ represents <u>future null</u> infinity. We expect, therefore, that zero rest-mass fields are to be significant at $\mathscr{I}^-$ and $\mathscr{I}^+$, but that fields of finite rest-mass are important at I^- and I^+ and not at $\mathscr{I}^-$ and $\mathscr{I}^+$.

Consider next the Schwarzschild metric:

$$(7.5) \qquad ds^2 = du^2\left(1 - \frac{2m}{r}\right) + 2drdu - r^2\left(d\theta^2 + \sin^2\theta\, d\varphi^2\right)$$

where one null (retarded) coordinate u has been used. Introducing $d\hat{s} = \Omega\, ds$ with $\Omega = r^{-1}$ and choosing a new coordinate $\ell = r^{-1}$ we get

$$(7.6) \qquad d\hat{s}^2 = du^2(\ell^2 - 2m\ell^3) - 2d\ell du - d\theta^2 - \sin^2\theta\, d\varphi^2.$$

This is regular (C^∞) at the hypersurface $\mathscr{I}^+$ defined by $\ell = 0$, u finite. Similarly by re-expressing the metric in terms of an advanced parameter v in place of the retarded parameter u, we get $\mathscr{I}^-$ as regular (C^∞). In each case $\mathscr{I}^+$, $\mathscr{I}^-$ will be topologically a three-dimensional open cylinder ($S^2 \times E^1$). This is precisely the same situation as in the Minkowski case if we exclude the points I^-, I°, I^+. However, if we wish to <u>include</u> these points in the manifold $\hat{\mathscr{M}}$ for the Schwarzschild case, a new situation arises. This is that each of I^-, I°, I^+ must apparently be necessarily a <u>singular</u> point for $\hat{\mathscr{M}}$. That I^- and I^+ are singular is not surprising since all the matter is concentrated at these points when it meets the boundary. However, it also seems that I° must

also be singular for any asymptotically flat case where the total mass is non-zero -- that is to say, in any physically interesting asymptotically flat case except for Minkowski space-time.

The form of Schwarzschild metric given above is a special case of

$$(7.7) \qquad ds^2 = r^{-2} A dr^2 - 2B_i dx^i\, dr + r^2 C_{ij} dx^i\, dx^j$$

$(i, j = 1, 2, 3)$, where if $r^{-1} = x^0$, then A, B_i, C_{ij} are to be sufficiently (say C^3) differentiable functions of x^μ ($\mu = 0, 1, 2, 3$) at, and in the neighbourhood of, the hypersurface $\mathscr{I}$ defined by $x^0 = 0$. Assuming, also, that the relevant determinant in A, B_i, C_{ij} does not vanish, this clearly means that the metric $d\hat{s} = \Omega ds$ with $\Omega = r^{-1}$ is regular (C^3) on $\mathscr{I}$. The metric (7.7) includes as particular cases many which have been suggested as representing gravitational radiation in asymptotically flat space-times [69]. The de Sitter metric also readily takes the form (7.7) (with $A = 1$, $B_i = 0$, $C_{ij} = \delta_{ij}$), where r is the exponential of the usual time parameter. Furthermore, some asymptotically de Sitter metrics which generalize those of Robinson-Trautman to include a cosmological term, also have the form (7.7). It would therefore seem that an assumption of regularity for $\mathscr{I}$ is not an unreasonable one to impose on a space-time if properties involving gravitational radiation are to be studied.

The choice of Ω for (7.7) has the important property that its gradient is non-vanishing ($\nabla_\mu \Omega \equiv \partial\Omega / \partial x^\mu = (1, 0, 0, 0)$ at $\mathscr{I}$) and it therefore determines the normal direction to $\mathscr{I}$ (since $\mathscr{I}$ is given by $\Omega = 0$). It is often useful to supplement an assumption of regularity for $\mathscr{I}$ with this condition that the gradient of Ω be non-vanishing there. However, this condition can sometimes be deduced from the regularity assumption if certain additional physical assumptions are made.

To be more precise about the conditions to be satisfied, first, for $\mathscr{M}$ to be a space-time, we require on physical grounds that:

(7.8a) $\mathcal{M}$ is a suitably differentiable, connected Riemannian 4-manifold without boundary, with metric ds of signature (+, -, -, -) in which 'past' and 'future' null cones can be distinguished consistently and which contains no closed smooth timelike curves.

Secondly, the space-time $\mathcal{M}$ will be called <u>asymptotically simple</u> if some manifold $\hat{\mathcal{M}}$ with boundary $\mathscr{I}$ ($\subset \hat{\mathcal{M}}$) and metric $d\hat{s}$ exists, whose interior $\hat{\mathcal{M}} - \mathscr{I}$ is conformal to $\mathcal{M}$ with $d\hat{s} = \Omega\, ds$ ($\Omega > 0$) and which satisfies the following properties:

(7.8b) This manifold $\hat{\mathcal{M}}$ is differentiable (say C^4) everywhere, with regular (say C^3) metric $d\hat{s}$. (Regularity at points of $\mathscr{I}$ is to be interpreted as the possibility of regarding $\hat{\mathcal{M}}$ as a part of a more extended four-dimensional manifold of which $\mathscr{I}$ is a set of regular interior points.)

(7.8c) Defining $\Omega = 0$ on $\mathscr{I}$, we have Ω regular (say C^3) throughout $\hat{\mathcal{M}}$. Furthermore, $\nabla_\mu \Omega \neq 0$ on $\mathscr{I}$.

(7.8d) Every null geodesic in $\hat{\mathcal{M}}$ contains, if maximally extended, two distinct points on $\mathscr{I}$.

The reason for condition (7.8d) is to ensure that $\mathscr{I}$ includes the <u>whole</u> of the 'null infinity' for $\mathcal{M}$. (A null geodesic remains a null geodesic after conformal transformation.) Condition (7.8d) is perhaps somewhat unnecessarily severe and a weakened 'local' alternative version of condition 4 could perhaps be given if desired, although its precise statement might be considerably more involved.

It is important to observe that the <u>conformal geometry</u> of $\hat{\mathcal{M}}$ is completely determined by the physical space-time $\mathcal{M}$. If $\hat{\mathcal{M}}$ and $\tilde{\mathcal{M}}$ are related to $\mathcal{M}$ via $d\hat{s} = \Omega\, ds$, $d\tilde{s} = \tilde{\Omega}\, ds$, respectively, then the <u>whole</u> of $\hat{\mathcal{M}}$ is related by a non-

singular conformal transformation $d\tilde{s} = \Omega' \, d\hat{s}$ ($\Omega' > 0$) to the whole of $\tilde{\mathcal{M}}$. For we have $\Omega' = \tilde{\Omega}\,\Omega^{-1}$ in the interior region and $\Omega' = \nabla_\mu \tilde{\Omega} : \nabla_\mu \Omega$ on the boundary. (The vectors are proportional since they are both normal to the hypersurface $\mathscr{I}$.) Thus, $\hat{\mathcal{M}}$ is physically meaningful if regarded as a conformal manifold. It is convenient for calculations however, to use a particular metric structure for $\hat{\mathcal{M}}$ and to carry out covariant differentiations with respect to this metric. It is not necessary that the particular metric chosen for a calculation be defined globally on $\hat{\mathcal{M}}$ if local conformal properties of $\mathcal{M}$, only, are being investigated.

If, in addition to asymptotic simplicity, certain physical assumptions are made about the nature of the energy-momentum tensor

$$(7.9) \qquad T_{\mu\nu} = -\frac{1}{2\kappa}\left(R_{\mu\nu} - \frac{1}{2} R\, g_{\mu\nu} + \lambda\, g_{\mu\nu}\right),$$

(cf. (4.18)), then a number of interesting consequences follow.[46] In particular, if we assume that $T_{\mu\nu}$ does not asymptotically approach a non-zero multiple of $g_{\mu\nu}$, then $\mathscr{I}$ is spacelike, timelike, or null according as λ is positive, negative, or zero. Furthermore, if $T_{\mu\nu} = 0$ near $\mathscr{I}$, then we can deduce the peeling-off property for the gravitational field,[71,61,46] namely, the fact that the Weyl tensor components (6.17) have the asymptotic behaviour (cf. (6.32), (6.52)) that they fall off as successive powers of r: r^{-5}, r^{-4}, r^{-3}, r^{-2}, r^{-1}. If the Einstein-Maxwell equations hold near $\mathscr{I}$, then a similar result can be obtained for both the gravitational and electromagnetic fields. (For electromagnetic "peeling-off" the field components fall off[71,72,46] as r^{-3}, r^{-2}, r^{-1}, cf. (6.32), (6.52).) In fact, a result of this kind will hold generally for zero rest-mass fields of arbitrary spin if we represent the field in a conformally covariant way (cf. (4.5), bearing in mind (4.17)). The condition that the field, when conformally transformed, remains finite at the boundary $\mathscr{I}$, is simply another way of stating the "peeling-off" property of the field.

When $\mathscr{I}$ is null, we may think of asymptotic simplicity as a definition of asymptotic flatness. When the space-time $\mathscr{M}$ is empty near $\mathscr{I}$, this is in very close agreement with the Bondi-Sachs[61] and Newman-Unti[60] definitions, while being completely coordinate independent. (It is possible, however, that in general when $\mathscr{M}$ is not empty near $\mathscr{I}$, we should add some further conditions.) Also, the topological nature of $\mathscr{I} = \mathscr{I}^+ \cup \mathscr{I}^-$, when $\mathscr{I}$ is null, namely $\mathscr{I}^+ \simeq \mathscr{I}^- \simeq S^2 \times E^1$, is another consequence of the asymptotic simplicity conditions. When $\mathscr{M}$ is spacelike (or timelike) we may think of asymptotic simplicity as implying that $\mathscr{M}$ is asymptotically de-Sitter.

We may specify initial data for zero rest-mass fields on $\mathscr{I}$. When $\mathscr{I}$ is null, the null-data (cf. section V) simply gives us the radiation fields. Thus, the use of incoming radiation field as initial data is closely related to the ideas of section V. When $\mathscr{I}$ is spacelike, however, we are presented with a problem: the radiation field concept turns out to be origin dependent.[46] This is because no unique null direction is given on $\mathscr{I}$ to define a unique component describing the $O(r^{-1})$ part of the field. This has the consequence that (for spin greater than one-half) it does not appear to be possible, for general source distributions, to define a purely retarded field when $\mathscr{I}^-$ (past infinity) is spacelike, or a purely advanced field when $\mathscr{I}^+$ (future infinity) is spacelike.[66] This is closely related to the question of visual horizons in cosmological models.

Although most cosmological models are not asymptotically simple, we can discuss them using the conformal technique[66] if we drop the condition (7.8c). For example,[73] in the case of the Einstein-de Sitter model we necessarily have $\nabla_\mu \Omega = 0$[74] at $\mathscr{I}^+$. We can also represent $\mathscr{I}^-$ for this model if we allow $\Omega = \infty$ at $\mathscr{I}^-$. This is an expanding model which starts from an initial point singularity. The initial singularity can still be represented as a non-singular boundary hypersurface of $\hat{\mathscr{M}}$. This is a kind of opposite procedure to

that considered earlier. Instead of the infinite compression ($\Omega = 0$) needed to make infinity finite, this time an infinite expansion ($\Omega = \infty$) turns a singular point into a non-singular hypersurface. This initial hypersurface turns out to be spacelike. On the other hand, $\mathscr{I}^+$ turns out to be null, as in the Minkowski case, with a future vertex I^+ and a past vertex I^0 that lies on $\mathscr{I}^-$ The world-lines of "galaxies" ("fundamental observers") are lines originating at points of $\mathscr{I}^-$ and terminating at I^+ (see Fig. 6).

In the steady-state model there is no initial explosion. However, in this case, $\mathscr{I}^-$ still does not represent a region properly at infinity. The model is not <u>complete</u> in the past in the sense that there are time-like geodesics which cannot be extended to infinite length into the past. Only <u>half</u> of a complete de Sitter hyperboloid (not to be confused with the Einstein-de Sitter model) is represented by the steady-state universe. The boundary $\mathscr{I}^-$ is an ordinary null hypersurface with respect to the <u>physical</u> metric. Thus we have Ω finite (non-zero) at $\mathscr{I}^-$. On the other hand, the model is asymptotically simple with respect to <u>future</u> directions: $\mathscr{I}^+$ is space-like with $\Omega = 0$, $\nabla_\mu \Omega \neq 0$. Also $\mathscr{I}^-$ has a past vertex I^- at which $\Omega = 0$ and a future vertex I^0 lying on $\mathscr{I}^+$. The conformal manifold $\hat{\mathscr{M}}$ with boundary is, in fact, identical with that for the Einstein-de Sitter model, except that the time direction is reversed (see Fig. 7). Again, this model represents an expanding universe. The world lines of galaxies are represented (because of the continual creation) as lines originating in the interior of $\hat{\mathscr{M}}$ and terminating on $\mathscr{I}^+$. There is an infinite density of these lines at $\mathscr{I}^+$ (<u>and</u> at $\mathscr{I}^-$) but not in the interior of $\hat{\mathscr{M}}$. Each such line if extended backwards would pass through I^- (assuming no proper motions).

As with the discussion of the Minkowski case, each of these models can be represented conformally as a portion of the Einstein static universe (Fig. 8).

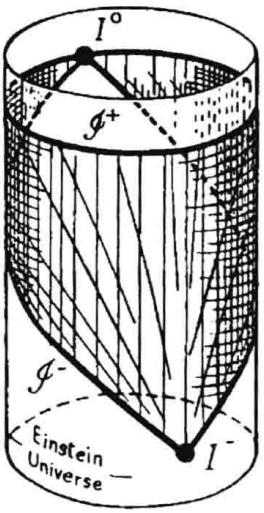


Fig. 8. The representation of the steady-state model conformally as part of an Einstein universe. This brings out that I^0 is a single point. The representation of the Einstein–de Sitter model is the same but upside-down. In the case of the relativistic models with $k < 0$, I^0 is not a point because $\mathscr{I}$ does not wrap all the way around the cylinder.

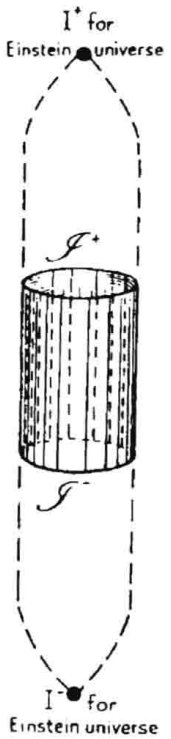


Fig. 9. The complete de Sitter space-time represented on the Einstein universe (shown dotted). This model can be obtained by joining two copies of the model shown in Fig. 8, one of which is inverted. Except for the cases which are asymptotically Einstein universes, every relativistic cosmology with $k > 0$ has a similar representation.

In this case, just half the region representing Minkowski space-time is covered. To represent the complete asymptotically simple de Sitter space-time (contracting and expanding model), two of these portions must be joined along their null boundary -- one being inverted. The result is simply a finite cylindrical section of the infinite cylinder representing the Einstein static universe (Fig. 9). The boundary surfaces $\mathscr{I}^-$ and $\mathscr{I}^+$ are disjoint and non-singular, each being a three-dimensional sphere (S^3). In fact, this particular representation applies also to all the other relativistic (vanishing pressure) Robertson-Walker models with positive spatial curvature k (e.g., the Lemaitre and Friedmann models), except that the cylinder becomes semi-infinite for the two models (e.g., Eddington-Lemaitre) which are asymptotically Einstein universes. The Einstein universe itself is a kind of maximal universal conformally flat space-time. That is to say, every conformally flat model is conformal to a portion of the Einstein universe (possibly with identifications). The infinity of the Einstein universe cannot be represented as a non-singular hypersurface. Any conformal factor which makes the "ends" of the cylinder finite must also pinch the cylinder to a point at each end. A compression in the time direction must be accompanied by an equal compression in the space directions. Thus, there is just an initial singular point I^- and a final singular point I^+ for this case.

The Einstein-de Sitter model is the simplest expanding relativistic (vanishing pressure) cosmology, since it has $k = 0$, $\lambda = 0$. The corresponding models with $k = 0$ and $\lambda < 0$ or $\lambda > 0$ have very similar finite conformal representations except that now both $\mathscr{I}^+$ and $\mathscr{I}^-$ are spacelike. They are joined to each other by the single point I^0. The cases with $k < 0$ correspond closely to those with $k = 0$. Their finite conformal representations differ only in that I^0 is now no longer a point but a spacelike spherical surface (S^2). The Milne model's representation differs from that of Minkowski space-time in precisely the same

way. When represented on the Einstein universe, unlike the cases with k = 0, these models do not wrap all the way around the cylinder.

In most physically interesting cases each of $\mathscr{I}^-$ and $\mathscr{I}^+$ is either spacelike or null (An example for which $\mathscr{I}$ is timelike is afforded by the de Sitter model with negative λ .) The spacelike nature of $\mathscr{I}^-$ or $\mathscr{I}^+$ in many of these models is closely related to the existence of one or other of the two types of visual horizon. Suppose first that $\mathscr{I}^-$ is spacelike. Now, the world-lines of particles or "fundamental observers" will not all meet $\mathscr{I}^-$ at the same point. (For if they did, the space-time outside the null cone of this point should not properly be considered part of the universe. $\mathscr{I}^-$ ought then to be redefined as either this null cone or possibly a time-like hypersurface within it.) Consider a particular observer and an event 0 on his world-line sufficiently close to $\mathscr{I}^-$. The past null cone of 0 will intercept some but not all of the particles in the universe before it meets $\mathscr{I}^-$ (see Fig. 10). As 0 moves into the future along the observer's world-line, more and more particles come into view. This division of the particles into those seen and those not seen is termed a particle horizon. On the other hand, if $\mathscr{I}^-$ is null, we expect that the world-lines of the fundamental observers all pass through a past vertex I^- of $\mathscr{I}^-$. The past null cone of any event 0 near I^- clearly intercepts all these world-lines and there is no particle horizon (Fig. 11.)

Suppose, now, that $\mathscr{I}^+$ is spacelike. The world-line of any observer terminates at point 0* on $\mathscr{I}^+$. The past null cone of 0* divides the universe into events which the observer can see at some stage of his history and those events that he can never see no matter how long he waits (Fig. 12). Suppose a clock crosses the past null cone of 0* at some event at which the clock registers 12:00. Then, the observer sees the clock register 11:30, 11:55, 11:59, ..., but he will never see it register 12:01. To an observer sitting on the clock,

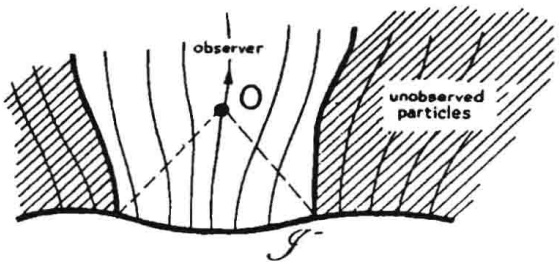


Fig. 10. When $\mathscr{I}^-$ is space like, a particle horizon occurs. The observer can see only some of the particles in the universe. As his time progresses, more and more particles come into view.

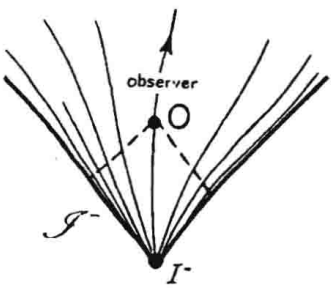


Fig. 11. No particle horizon occurs if $\mathscr{I}^-$ is null since all particles can always be seen by the observer (assuming no continual creation).

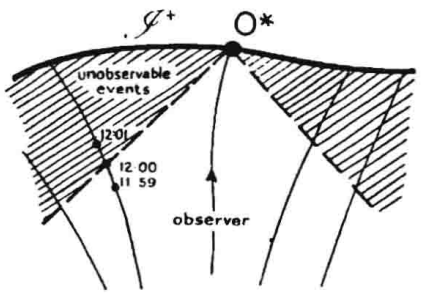


Fig. 12. An event horizon occurs when $\mathscr{I}^+$ is space-like. The events outside the past null cone of 0* are never seen by the observer.

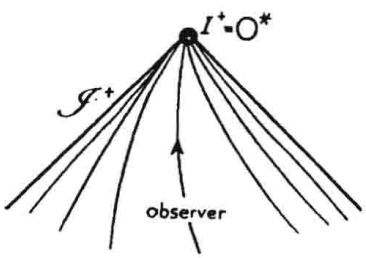


Fig. 13. No event horizon arises if $\mathscr{I}^+$ is null since the past null cone of 0* coincides with $\mathscr{I}^+$.

however, nothing special occurs at 12:00. This division of events into those ultimately seen and those never seen is termed an event horizon. If $\mathscr{I}^+$ is null, we expect the world-lines of the fundamental observers all to pass through a future vertex I^+ of $\mathscr{I}^+$. In this case $0^* = I^+$ and its past null cone is the future boundary to the universe. Thus, no event horizon occurs (Fig. 13). We, therefore, have the result that:

(7.10a) a particle horizon corresponds to $\mathscr{I}^-$ spacelike,

(7.10b) an event horizon corresponds to $\mathscr{I}^+$ spacelike.

(For a further discussion, which connects these features with the question of advanced and retarded fields in cosmological backgrounds, see ref. 66.)

VIII. COLLAPSE AND COSMOLOGY

A question of great interest in relativity theory and cosmology is that of the existence of singularities. The usual "big bang" models of relativistic cosmology are characterized by the presence of an initial state of infinite curvature and infinite density. At such an initial singularity the ordinary view of space-time as a smooth manifold would have to break down. Also, the existence of a particle horizon presents us with an essential problem, since different highly curved portions of the universe, which previously could not communicate with one another, must come together in such a way that they fit consistently. The only previous "causal" link between these portions was the initial singularity itself. We are thus presented with a real problem of principle in addition to all the formidable, but essentially technical problems, concerned with the physics of early high density states of these models.

It is a reasonable question to ask, however, whether the initial singularity of these models is not more a consequence of the mathematical idealizations involved, than of the actual physical situation that the model is supposed to describe. In particular, it is virtually essential to postulate a high degree of symmetry (e.g. spatial homogeneity), for an exact model to be amenable to detailed mathematical treatment. And with such symmetry present, the picture is almost inescapable that all the matter in the universe should, at some time, have been compressed simultaneously into a single point, or at least into a line or into a two-dimensional surface.[75] Possible ways out of the dilemma, that have been suggested, include the adoption of a large enough cosmological constant, the presence of "unreasonable" matter (e.g. negative energy densities -- models with continual creation of matter would have to come under this heading), sufficient rotation (e.g. as in Gödel's model), or a change in the laws of physics which could have been operative at very early stages of the universe.

In addition, there are possible quantum effects of gravitation to be considered, although these would apparently only be important at "ridiculously" high densities[17] (e.g. 10^{93} g/cc). But more significant in the first instance than any of these other considerations, to my mind, is the role played by local irregularities of matter and curvature. In fact, near the singularity of the model, we really have no reason at all to expect that it should be realistic. The situation is perhaps clearer for the time-reversed process. If we imagine all the matter in the universe hurtling simultaneously to a single "central point" (or "line," etc.) then an infinite density at that point (or line, etc.) is hardly surprising. If, however, we perturb slightly the motion of the incoming matter, so that it is not exactly focused to one point (line, etc.) then we might expect that although the density could become very high, it would not become infinite. The possibility then may be considered of an effective "bounce" to the universe. Thus the presently observed expanding state of the universe might be regarded as the result of a previously contracting phase, the transition from contraction to expansion being achieved via a condensed but highly asymmetrical and complicated intermediate phase.

A picture of this kind has been suggested by many authors.[76] Also some detailed considerations by Lifshitz and Khalatnikov[77] had suggested that such avoidance of singularities by small perturbations of the motions might indeed be possible. However, recent work has ruled out many possibilities of this kind. In particular, Hawking[78] has shown that a spatially closed universe which is collapsing at every point of a space-like cross-section cannot "bounce" without encountering space-time singularities, that is, unless a certain "physically reasonable" inequality on the Ricci tensor is violated. I shall discuss this and another of Hawking's results shortly. In the meantime I shall give a new version of an earlier result of my own[79] which applies to the closely related

phenomenon of the singularities encountered in gravitational collapse. Some possible cosmological implications of these considerations will be given at the end.

The essential similarity between gravitational collapse and the time-reverse of the early stages of a "big bang" cosmology has been emphasized particularly by Wheeler.[80] Indeed, there are certain advantages in concentrating attention on the more local phenomenon of collapse, rather than on the global cosmological question. For we do not need to know the entire large-scale structure of the universe to study a local phenomenon (or do we?); if the local laws of physics are strongly affected by the structure of the universe as a whole ("Mach's principle"), then we might at least expect the local phenomenon to be the less sensitive to such changes in these laws; finally, unlike the universe as a whole, collapse is not something given to us just once, so we are freed to exploit the immense theoretical potentialities of the "gedanken experiment". My own personal feeling is that the deeper aspects of the cosmological problem will not be solved without a simultaneously deeper appreciation of what is involved in collapse.

But what is meant, precisely, by "gravitational collapse"? The idea comes originally from the classical study large spherically symmetrical bodies. According to the work of Chandrasekhar,[81] a cold spherically symmetrical (non-rotating) body, which has reached the endpoint of thermonuclear activity, cannot hold itself apart against gravitational forces if its mass is much greater than that of the sun. Such a body would therefore collapse and there would be nothing to halt this collapse. Later analysis has confirmed this conclusion.[82] What then would be the fate of such a collapsing star? Assuming exact spherical symmetry, the matter would fall radially inwards through its Schwarzschild radius r = 2m, to encounter a space-time singularity at the centre r=0. An

observer situated at large distances from the star would never see the collapse to within r = 2m, however. To him, the star's collapse would appear to slow down and approach r = 2m asymptotically.

Before considering how physically realistic this picture is, let us examine the solution of Einstein's equations on which it is based, namely the Schwarzschild solution. This defines the spherically symmetrical gravitational field outside the collapsing star. We have, with the usual Schwarzschild coordinates:

$$(8.1) \qquad ds^2 = (1-2m/r)dt^2 - (1-2m/r)^{-1} dr^2 - r^2 (d\theta^2 + \sin^2\theta \, d\varphi^2).$$

Let the star's boundary be given by r = f(t). The metric (8.1) then only applies outside this boundary, i.e. $r \geqslant f(t)$ (with the usual inequalities and identifications holding for θ and φ). The star's boundary must be timelike: $|f'(t)| \leqslant 1-2m$, from which it follows that this boundary never crosses r = 2m in (8.1) (fig. 14). One might be inclined to infer that the star therefore always remained outside r = 2m. But if we calculate the total proper time of a particle on the surface of the star, where we assume, for simplicity, that the star falls freely inwards, we find that this proper time is finite. Thus, an observer who follows the star inwards must, after this finite time has expired, either of necessity be destroyed (e.g. because he encounters infinite tidal forces or some other form of singularity) or else, find himself in a portion of the universe not covered by the coordinates of (8.1).

In fact, in this instance it is the latter possibility which holds. We can see this if we transform the coordinates in (8.1) by introducing an advanced time parameter $v = t + r + 2m \log (r - 2m)$. (The question of uniqueness of this process of "analytic extension" will be considered shortly.) Then the metric takes the Eddington-Finkelstein [83] form (compare 7.5)

$$(8.2) \qquad ds^2 = (1 - 2m/r)\, dv^2 - 2\, dv\, dr - r^2(d\theta^2 + \sin^2\theta\, d\varphi^2).$$

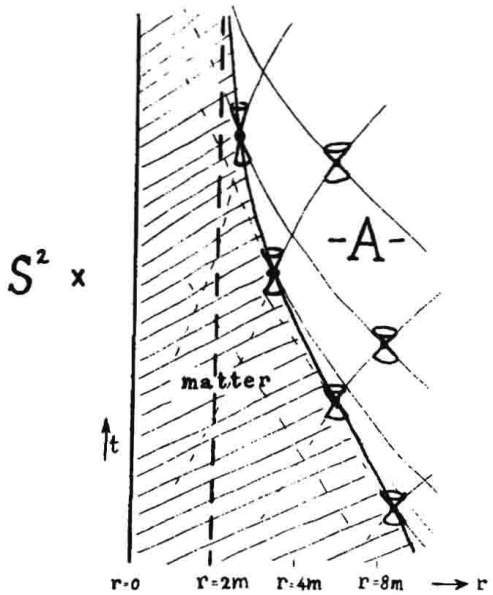


fig. 14: Schwarzschild picture

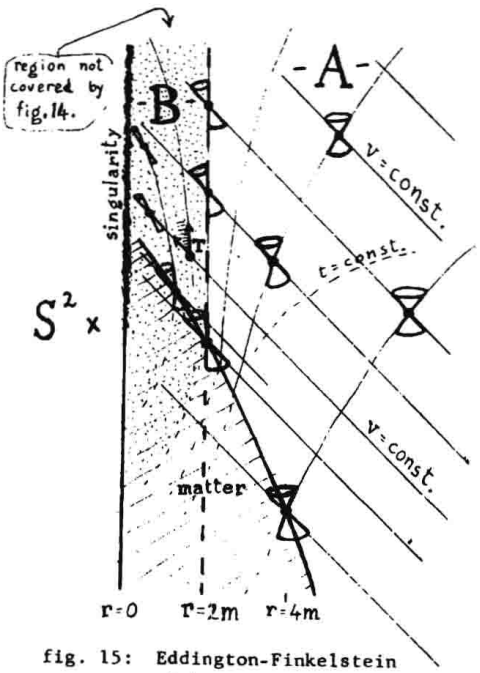


fig. 15: Eddington-Finkelstein picture

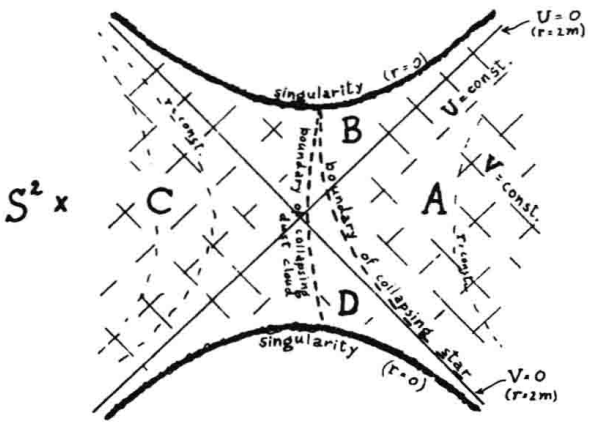


fig. 16: Kruskal picture.

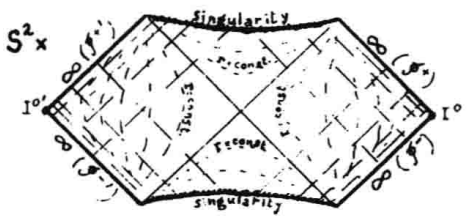


fig. 17: Kruskal with infinity represented.

When $r > 2m$, the forms (8.1) and (8.2) are precisely equivalent, but (8.2) has the advantage that the metric covers a larger region in a non-singular way. The metric in the neighbourhood of the (null) hypersurface $r = 2m$ is perfectly regular. It is only because the more familiar form (8.1) possesses a coordinate singularity at $r = 2m$, that this hypersurface is sometimes referred to as the "Schwarzschild singularity." The surface of the star crosses $r = 2m$ at a definite value of v (although, of course, the here unsuitable t parameter becomes infinite) and proceeds inwards towards $r = 0$. But because of the positioning of the light cones (see fig. 15), an outside observer cannot see any of the empty region $r \leqslant 2m$. More particularly, this is because, in the empty region, $r = 2m$ is a null hypersurface which lies "above" (i.e. to the future of) the entire external region. An observer or signal can cross the hypersurface from the outside to the inside but not the other way around. Because of the analogy with the similar situations in cosmology, this hypersurface is frequently referred to as an event horizon.

If we trace the history of the star's surface inside the $r = 2m$ region, we see that it must, of necessity, encounter $r = 0$. For the surface must continue to move in a timelike direction and the null-cones are tipping over more and more towards the $r = 0$ axis. Again if we calculate the total proper time of a particle on the surface of the star (assuming free fall or otherwise), up until the point at which it encounters $r = 0$, we find that this is finite. But this time there is no hope of extending the solution further. If we evaluate the curvature scalars constructed from the Weyl tensor, we find that they tend to infinity as r approaches zero. Thus, our observer who successfully followed the star through $r = 2m$ must now be torn to pieces by the infinite tidal forces at the true singularity at $r = 0$.

This is the picture presented by Oppenheimer and Synder [84] when they considered the dynamics of a collapsing uniform dust cloud. In fact the space-time interior to the dust cloud can also be described simply and explicitly and turns out to be nothing other than (a portion of) the Friedmann universe.[73] This again emphasizes the essentially identical nature of the situation of a final singularity in gravitational collapse and of a final (or initial) singularity for a relativistic cosmology.

If we wish to consider the entire history of the Oppenheimer-Synder gas cloud and not just the final collapse phase, then even the metric form (8.2) does not cover a sufficient portion of the empty region. The behaviour of the dust cloud is time-symmetrical: like the Friedmann universe, it starts from an initial singularity, expands to a maximum volume and then recontracts to a final singularity. Thus we may expect to join the matter-filled region to an empty, spherically symmetrical exterior region which is also time-symmetrical. We can, in fact, exhibit such a region if we use the U, V coordinates of Kruskal,[85] which are related to the Schwarzschild r, t (in the region $r > 2m$) by

(8.3a) $$V/U = - e^{t/2m}$$

(8.3b) $$VU = e^{r/2m} (1 - r/2m).$$

The metric then takes the form

(9.4a) $$ds^2 = f^2 \, dU \, dV - r^2(d\theta^2 + \sin^2\theta \, d\varphi^2)$$

where

(8.4b) $$f^2 = (32 \, m^3/r) \, e^{-r/2m} .$$

The matter region is joined to the Kruskal region along a timelike hypersurface. This hypersurface is described by a timelike geodesic in the Kruskal (U, V)-plane (fig. 16). Thus, only the portion to the right of this geodesic in the diagram is to be considered in this model. The left-hand portion of fig.16

must be replaced by a spherically symmetric portion of the Friedmann universe (i.e. given by $0 \leqslant \chi \leqslant \chi_0$ in $ds^2 = A^2 (1 + \cos\eta)^2 (d\eta^2 - d\chi^2 - \sin^2\chi (d\theta^2 + \sin^2\theta\, d\varphi^2)$; A, χ_0 being constants). The boundary is again a hypersurface generated by timelike geodesics (now the geodesic $\chi = \chi_0$ in the (η, χ)-plane). These two portions do, in fact, fit together adequately smoothly.[86]

One drawback of the Kruskal form of the Schwarzschild metric is that (8.4) is not given "explicitly", depending on the solution, for r, of (8.3b). The Eddington-Finkelstein form (8.2) is, however, much simpler, and it is interesting that the main features of the Kruskal extension can be obtained simply[87] from (8.2). To achieve this, we note that (8.2) covers the two portions A, B of the Kruskal diagram (fig. 16) (with $2mV = e^{v/2m}$). By using a retarded parameter u in place of the advanced parameter v, we can likewise cover the portions A, D of fig. 16 with another "coordinate patch", the overlap region being A. In an exactly similar way, we can cover the portion C, B or the portion C, D with a "coordinate patch" having the form of (8.2). The coordinate transformations in the overlap regions C and D can be obtained, if desired, by transforming back to the form (8.1) (with $0 < r < 2m$) as an intermediate stage. The one feature of the Kruskal extension which is not obtained by this method is the regular nature of the space-time at the 2-sphere $U = V = 0$. However, the method has a considerable heuristic value in that it applies at once to certain other more complicated extensions, notably those of the Kerr and Reissner-Nordström solutions (cf. (8.27), (8.28)).[88] In this connection it will be useful to have, also, the version of the Kruskal diagram in which null infinity is represented in terms of finite hypersurfaces in the manner of section VII (e.g. use coordinates $p = \tan^{-1} \sinh^{-1} V$, $q = \tan^{-1} \sinh^{-1} U$; see fig. 17).

Let us return, now, to the case of the collapsing star. Can the description that I have given be regarded as at all realistic? For example, even if we

retain the exact spherical symmetry, can it be said that we really know enough about the properties of matter at the density (approximately the same as that of an atomic neucleus, for the case of an ordinary star) at which the star is supposed to fall through its "Schwarzschild throat"? And might it not be that a massive star inevitably ejects sufficient material, as it nears r = 2m, that its mass invariably falls below Chandrasekhar's limit, thus enabling a stable self-supporting state to become possible? It seems that for many years, the prevailing belief had been that a collapsing star would always be able to save itself by such means. But apparently this was largely because astronomers did not feel the need to consider self-supporting masses that were enormously greater than that of the sun. But with the discovery of the quasi-stellar objects,[89] renewed interest in the subject of gravitational collapse has been stimulated. For, considering the enormous quantities of energy that these objects appear to be emitting, and their remarkably small size, it has been suggested that "individual" masses of the order of 10^6 or 10^8 times that of the sun may be involved. Now the Schwarzschild radius of a spherically symmetric body is proportional to its mass. Thus the chracteristic density at which such a body crosses r = 2m should be inversely proportional to the square of its mass. For an object of 10^6 or 10^8 solar masses, this density is not unreasonably high. Indeed, Fowler has emphasized that for an object of 10^{11} solar masses (such as a good sized galaxy) this characteristic density is less than that of air! And there seems no reason at all to believe that such an object would necessarily eject virtually all its material before reaching r = 2m.

Then what about the role played by deviations from spherical symmetry? For example, can one even <u>define</u> the analogue of r = 2m for an asymetrical body? And might it not be that the presence of <u>rotation</u> always prevents the final collapse? The question of rotation is a particularly pertinent one.

Indeed, Ostriker, Bodenheimer and Lynden-Bell [90] have recently shown that if (differential) rotation is present, then equilibrium states do exist for objects which can be of mass very much larger than Chandrasekhar's limit (arbitrarily large, in fact, within the essentially Newtonian framework used). Now, fortunately, there is a known solution of Einstein's vacuum equations, namely that of Kerr, [91, 88] which generalizes the Schwarzschild solution by the inclusion of angular momentum. The Kerr solution contains two arbitrary parameters m and a, where m defines the mass as in Schwarzschild's solution, and where ma defines the angular momentum. The solution (given explicitly in (8.27)) is still of a rather special character (for example, the quadrupole moment has to be given by $2\,ma^2$) but it seems likely that at least the main features of the geometry of the space-time surrounding a rotating and possibly collapsing body can be gleaned from a study of Kerr's solution. One of the most significant features of the solution, in fact, is that if $a > m$ there is no analogue of the "Schwarzschild throat", that is, there is no event horizon which prevents signals from the interior regions being received outside. On the other hand, if $a < m$, there is an event horizon and the solution qualitatively resembles the Schwarzschild solution in that it possesses a similar "throat" which can finally swallow up the material of the (super-) star. There are other features of the solution which qualitatively differ, however, from the case of the Schwarzschild solution. These will be discussed shortly, but in the meantime, I shall concentrate on the question of the "throat" itself and give a definition which may be used as a criterion for its existence.

Let us return to the Schwarzschild solution. For convenience, consider the metric in the form (8.2) and examine fig. 15. We wish to single out a property of the region B which characterizes it as "peculiar" in some sense. It should be emphasized, however, that there is nothing locally peculiar about the B-region.

A small neighbourhood of any point in the B-region is as good a solution of Einstein's vacuum equations as any other, and there is nothing at all "singular" about it. But it seems that in some way the whole B-region is "shrinking" in an apparently inevitable way to the space-time singularity at $r = 0$. To express this "shrinkage" we need to single out a partly global aspect of the B-region -- but it must not be too global because we require a property that will not be destroyed by small perturbations in the metric at the "time" of collapse. Consider, then a point T in the B-region of fig. 15 which can lie just inside $r = 2$ (but outside the matter region). Such a point describes a spacelike 2-sphere $\mathcal{T}$ in the whole space-time, of total surface area $4\pi r^2$ $(r < 2m)$. Now any system of material particles whose world-lines intersect $\mathcal{T}$ must have a subsequent motion which is bounded by the velocity of light. That is to say, if we consider the extreme situation of a flash of (idealized) light emitted at $\mathcal{T}$, there will be an "outgoing" flash and an "ingoing" flash, described by the two null geodesics in fig. 15 which pass through T. These two null geodesics describe null hypersurfaces through $\mathcal{T}$ which together constitute the boundary of the region containing all permissable particle world-lines originating at $\mathcal{T}$. Now the point is that both parts of the boundary are "shrinking". Not only is the total surface area of the sphere described by a point moving up the "ingoing" null geodesic decreasing, but so also is the total surface area correspondingly decreasing for the "outgoing" flash. This, then, is the property of $\mathcal{T}$ which I shall use to characterize it as a trapped surface.[79] The presence of a trapped surface in a region will then be our indication that there may be something "peculiar" about the subsequent history of the region.

To be more precise, I define:

(8.5) a trapped surface $\mathcal{T}$ as a compact, spacelike 2-surface with the property that the null geodesics which meet $\mathcal{T}$ orthogonally locally converge in future directions.

The convergence of a system of null geodesics can be measured by the quantity ρ defined in section VI, where the ℓ^μ are tangent vectors to the geodesics (the m^μ being complex, null and orthogonal to ℓ^μ; $m_\mu m^\mu = -1$):

$$\rho = m^\mu \bar{m}^\nu \nabla_\nu \ell_\mu \,. \tag{8.6}$$

For a general congruence of null geodesics, ρ will have an imaginary part which measures the rotation of the congruence. The vanishing of this imaginary part is, then, the condition for hypersurface orthogonality of the congruence; equivalently, it is the condition for the geodesics to be generators of a (one-parameter) system of null hypersurfaces. When ρ is real, the value of ρ pertains only to the null geodesics in each individual hypersurface and not to the relationship between the null geodesics in neighbouring hypersurfaces. Thus, we can use ρ to describe the geometry of a null hypersurface. We have, in fact,[71,92]

$$D(a^{1/2}) = -\rho a^{1/2} \tag{8.7}$$

where, as in section VI, D denotes differentiation in the direction of ℓ^μ:

$$D \equiv \ell^\mu \nabla_\mu \tag{8.8}$$

and where a denotes the area of cross-section of an infinitesimal bundle of null rays lying in the chosen null hypersurface. Thus, the characteristic property of a trapped surface is that for each of the two null hypersurfaces passing through $\mathcal{T}$ (these being generated by the two systems of null geodesics meeting $\mathcal{T}$ orthogonally), the value of ρ is positive at each point of $\mathcal{T}$. As we move $\mathcal{T}$ up either one of these null hypersurfaces, its surface area decreases initially at every point $\mathcal{T}$.

For future reference, I shall also require the quantity analogous to ρ (section VI), which measures the shear of a congruence of null geodesics (with tangent vectors ℓ^μ):

(8.9) $\sigma = m^\mu m^\nu \nabla_\nu l_\mu$

The geometrical interpretation [71,92] of σ is obtained if we consider a small circle in a plane orthogonal to l^μ. If we "drag" the circle up the null geodesics of the congruence, the circle may become distorted into an ellipse. The magnitude of this distortion is measured by $|\sigma|$ and the direction of the minor axis (in relation to $\mathrm{Re}(m^\mu)$), by $\arg \sigma$. As with ρ, if ρ is real, σ pertains only to the structure of the individual null hypersurfaces and not to their relationship to each other. Thus σ can be used to describe the geometry of a null hypersurface. (The following way of representing the geometrical meanings of ρ and σ may be mentioned in passing.[44] Let c be a null geodesic of the system. Choose a plane element Π orthogonal to c at each point of c. If we use an infinitesimal complex number Z for "Argand coordinates" on Π, relating the coordinates on the different Π's by parallel transport, then a null geodesic of the system, neighbouring to c which intersects one -- and therefore all -- of the Π's, is governed by the equation $DZ = -\rho Z - \sigma \bar{Z}$. In particular, (8.7) follows from this.)

The idea, here, is to establish a connection between the existence of a trapped surface in a space-time and the subsequent development either of a singularity in space-time (such as r = 0 in (8.2)) which may involve infinite curvature, or else of some other space-time behaviour which would be very strange from the physical point of view. But before considering the rigorous mathematical result, let us ask the question whether it is reasonable to expect trapped surfaces to develop at all in our actual universe. For any individual case of a collapsing star or superstar, this is clearly a matter of the detailed astrophysics involved. The amount of rotation present, the mass loss through neutrino radiation, through gravitational radiation, the exact nature of the asymmetry present, the nature of the equations of state, magnetic fields, etc.; all these

may be relevant. But the essential point I want to make here is that there can be no reason of principle against a trapped surface developing. To illustrate this, I shall consider the following "gedanken experiment."

A technically advanced (but presumably exceedingly foolhardy) race of beings inhabit a galaxy (preferably elliptical) which contains about 10^{11} stars. By means of rockets, these beings contrive to alter the velocities of the stars, so that there remains practically no transverse component to the stars' motions. Furthermore, their radial motions are adjusted so that all the stars will fall towards the centre so as to reach its vicinity at almost the same time. (The mass-energy expended in this operation would be small compared to the mass of the stars.) The region they have to aim for has roughly a radius 2.5 times that of the solar system. There is plenty of room in a volume of this size for all the stars, and they can arrive there before any serious problem of stellar collisions arises. (They can even steer to avoid collisions, if necessary!) But does this result in the development of a trapped surface? The fact that it does, depends on the most primitive considerations (almost observational ones) of general relativity. We consider a flash of (idealized) light emitted at the centre at about the time the stars enter the critical region. Because of the focusing effect on idealized light rays [93] (i.e. on null geodesics) by the gravitational field of each star, the light rays which started off by diverging from the centre will all pass through or near to enough stars so that they start converging again. Under these circumstances it can be shown that a topological sphere which lies just inside the outermost boundary of the light rays will indeed be a trapped surface.

This example serves to illustrate, also, the fact that there is no reason why an observer should be "destroyed" as he enters the critical region. The

curvature is still extremely small in the neighbourhood of the trapped surface and the space-time there is perfectly regular.[94] Indeed, locally there is nothing peculiar about the trapped surface itself. Even in flat Minkowski space-time, many surfaces exist which are "locally trapped". That is to say, they satisfy the conditions of (8.5) except for the fact that they are not compact. To construct such a surface, we need only consider the intersection of the past null cones of two spatially separated events.

If, then, it is accepted that trapped surfaces can exist, what of the consequences? I shall show, in fact, that if we assume (provisionally) that the universe is open (i.e. not spatially compact), then our present knowledge of the laws of physics is insufficient to enable us to calculate even in principle the future history of such a collapsing system. In order to be more precise as to the mathematical statement involved, let us consider the situations in which one can predict the behaviour of a physical system, the appropriate laws of physics (assumed to be locally Lorentz covariant partial differential equations of suitable type, e.g. an exact set of section IV) being known. Suppose we are given data on an initial hypersurface, which are locally sufficient for the time-development of the fields in question. Under what circumstances will the data also be globally sufficient?

There are, in fact, two distinct ways in which locally adequate data can fail to be globally adequate. The first applies only to non-linear fields and is simply that singularities may arise in the fields, thus preventing further development of the equations. The second reason is, perhaps, more subtle and applies both to linear and to non-linear fields.[55] I shall not enter into the matter in full generality here, but the main features can be illustrated by an example. Consider Minkowski space-time and two alternative spacelike hypersurfaces $\mathcal{H}_1$ and $\mathcal{H}_2$ imbedded in it. With the usual Minkowski coordinates, $\mathcal{H}_1$

CHAPTER 28

is to be the hyperplane $t = 0$, while $\mathscr{H}_2$ is the "semi-hyperboloid" $t = -(1 + x^2 + y^2 + z^2)^{1/2}$. Now suppose we specify initial data on the two hypersurfaces, where, for simplicity we consider a linear theory such as Maxwell's free-space equations, or the Schrödinger-Klein-Gordon equation. In the case of $\mathscr{H}_1$ it is clear that the locally adequate Cauchy data will also be adequate to define the fields throughout the whole space-time. But this is not so in the case of $\mathscr{H}_2$. In fact, only in that part of the space-time given by $t \leqslant -(x^2 + y^2 + z^2)^{1/2}$ will the fields be determined uniquely. This is particularly clear in the case of Maxwell's equations, since any event which lies "above" this region can be intercepted by a plane (pulse) wave which vanishes over the whole of $\mathscr{H}_2$. What, then, is wrong with the hypersurface $\mathscr{H}_2$? In fact, nothing is intrinsically wrong with $\mathscr{H}_2$ itself (it is, in fact, intrinsically geodesically complete, for example), nor even with its neighbourhood in the space-time. What is wrong is that a global condition concerning the relationship of the hypersurface to the entire space-time fails to be satisfied, so information can reach an event in the space-time without ever having "registered" on $\mathscr{H}_2$. We can think of such information as being transmitted along a non-spacelike curve. Here, there are non-spacelike curves in the region to the future of $\mathscr{H}_2$ which come in from infinity without intersecting $\mathscr{H}_2$.

Let us consider then, in general terms what this condition amounts to. The condition will be required for Theorem (8.22) which is the main result of this section. Some other mathematical preliminaries will also be given. For convenience, I shall assume that the space-time $\mathscr{M}$ satisfies the "physically reasonable" causality conditions (2.3), (2.4), (2.8). Let $\mathscr{H}$ be any subset of $\mathscr{M}$ I shall call $\mathscr{H}$ <u>semi-spacelike</u> if no timelike curve in $\mathscr{M}$ contains more than one event in $\mathscr{H}$. The definition of a <u>global Cauchy hypersurface</u> (GCH) for $\mathscr{M}$ is then as follows:

(8.10) $\mathcal{H}$ is a GCH for $\mathcal{M}$ means $\mathcal{H}$ is semi-spacelike and if c is a closed subset of a non-spacelike curve in $\mathcal{M}$ then c is part of a non-spacelike curve meeting $\mathcal{H}$.

Now any GCH for $\mathcal{M}$ must be a closed subset of $\mathcal{M}$. For, let the sequence (X_n) converge to X, with $X_n \in \mathcal{H}$. Any maximal timelike curve through X must intersect $\mathcal{H}$, say at Y. But if $Y \neq X$, either $I_+(Y)$ or $I_-(Y)$ is an open set containing X (cf. (2.7)) and therefore some of the X_n. This is impossible since $\mathcal{H}$ is semi-spacelike. Thus $X \in \mathcal{H}$.

Another criterion for a GCH is:

(8.11) $\mathcal{H}$ is a GCH for $\mathcal{M}$ if and only if $\mathcal{H}$ is a semi-spacelike subset of $\mathcal{M}$ which intersects every maximal null geodesic of $\mathcal{M}$ in a non-vacuous compact set.

The proof of (8.11), which will be given here only when $\mathcal{H}$ is a spacelike hypersurface, however, depends on the notions of the domains of dependence $D_+(\mathcal{S})$, $D_-(\mathcal{S})$ and of the Cauchy horizons[95] $H_+(\mathcal{S})$, $H_-(\mathcal{S})$ for an arbitrary closed subset $\mathcal{S}$ of $\mathcal{M}$:

(8.12a) $D_+(\mathcal{S})$ is the compliment of the union of all closed-set timelike curves. (i.e. closed subsets of timelike curves) which are inextendible in past directions and which do not intersect $\mathcal{S}$.

Another way of putting this is: $X \in D_+(\mathcal{S})$ if and only if every closed-set timelike curve through X can be extended into the past until it meets $\mathcal{S}$. From (8.12a), $D_+(\mathcal{S})$ is a closed set. The definition of $D_-(\mathcal{S})$ is simply:

(8.12b) ...the time-reverse of (8.12a).

(8.13a) $$H_+(\mathcal{S}) = \{X: X \in D_+(\mathcal{S}),\ I_+(X) \frown D_+(\mathcal{S}) = \varnothing\},$$

(8.13b) $$H_-(\mathcal{S}) = \{X: X \in D_-(\mathcal{S}),\ I_-(X) \frown D_-(\mathcal{S}) = \varnothing\}.$$

CHAPTER 28

Thus, $H_+(\mathcal{S})$ is the "future boundary" of $D_+(\mathcal{S})$ and $H_-(\mathcal{S})$, the "past boundary" of $D_-(\mathcal{S})$. It follows at once from (8.13), that $H_+(\mathcal{S})$ and $H_-(\mathcal{S})$ are semi-spacelike. Furthermore, they are closed sets.

I shall be interested in studying the structure of Cauchy horizons and of a certain more general type of set which I shall refer to as a <u>semi-spacelike boundary</u> (SSB). Using the obvious notations

$$(8.14a)\quad I_+[\mathcal{K}] = \bigcup\{I_+(X) : X \in \mathcal{K}\}$$

$$(8.14b)\quad I_-[\mathcal{K}] = \bigcup\{I_-(X) : X \in \mathcal{K}\}$$

for the "future" and "past" of an arbitrary subset $\mathcal{K}$ of $\mathcal{M}$, the defining property of a SSB is:

(8.15a) $\mathcal{S}$ is a SSB <u>if and only if</u> for some set $\mathcal{K} \subset \mathcal{M}$,

$$\mathcal{S} = \{X : I_+(X) \subset I_+[\mathcal{K}],\ X \notin I_+[\mathcal{K}]\}.$$

This amounts to saying that a SSB is the (past) boundary of a set which is a union of sets $I_+(X)$. By taking $\mathcal{L}$ as the compliment, in $\mathcal{M}$, of $\mathcal{S} \cup I_+[\mathcal{K}]$, or conversely, $\mathcal{K}$ as the compliment of $\mathcal{S} \cup I_-[\mathcal{K}]$, we obtain:

(8.15b) $\mathcal{S}$ is a SSB <u>if and only if</u> for some set $\mathcal{L} \subset \mathcal{M}$

$$\mathcal{S} = \{X : I_-(X) \subset I_-[\mathcal{L}],\ X \notin I_-[\mathcal{L}]\}.$$

Thus, the definition of a SSB is time-symmetrical: a SSB is also the (future) boundary of a set which is a union of sets $I_-(X)$.

If $\mathcal{U}$ is a closed subset of $\mathcal{M}$, then the set

$$(8.16a)\quad I_-[\mathcal{U}] \cup D_+(\mathcal{U}) - H_+(\mathcal{U})$$

is the union of the sets $I_-(X)$ with $X \in D_+(\mathcal{U})$. The boundary of (8.16a) is therefore an example of a SSB. This boundary contains $H_+(\mathcal{U})$ as a part.

Similarly, the boundary of the time-reverse (8.16b) of (8.16a) is a SSB and contains $H_-(\mathcal{U})$ as a part. Thus, we can obtain information about the structure of Cauchy horizons, from the general structure of a SSB.

Let $\mathcal{S}$ be a SSB. Then $\mathcal{S}$ consists of four disjoint parts (some of which may be vacuous) $\mathcal{S}_N$, $\mathcal{S}_+$, $\mathcal{S}_-$ and $\mathcal{S}_0$, defined as follows: if $X \in \mathcal{S}$, then there <u>may</u> or <u>may not</u> exist events $Y, Z \in \mathcal{S}$ for which

(8.17a) $$I_+(X) \subset I_+(Y),\ I_-(X) \subset I_-(Z);$$

the different possibilities define the subsets $\mathcal{S}_N$, $\mathcal{S}_+$, $\mathcal{S}_-$, $\mathcal{S}_0$ according to the scheme

(8.17b)

$X \in$	$(\exists Z)$	$(\nexists Z)$	
	$\mathcal{S}_N$	$\mathcal{S}_+$	$(\exists Y)$
	$\mathcal{S}_-$	$\mathcal{S}_0$	$(\nexists Y)$

The intuitive meanings of these subsets is that $\mathcal{S}_N$ represents the portion of $\mathcal{S}$ which is <u>null</u>, where $\mathcal{S}_+$ and $\mathcal{S}_-$ consist, respectively, of the future and past end-points of $\mathcal{S}_N$; $\mathcal{S}_0$ represents the spacelike portion of $\mathcal{S}$ (see fig. 18).

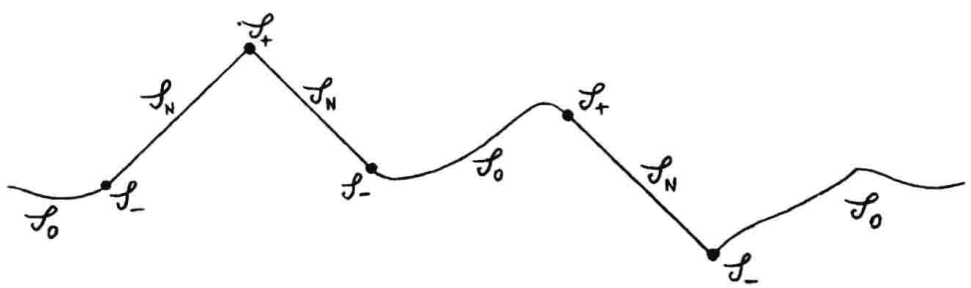


fig. 18

The different portions of a semi-spacelike boundary. (Null directions are inclined at 45°.)

A more precise statement is contained in the following lemma, which also gives a somewhat more useful condition for an event to lie in $\mathcal{S}_N$, $\mathcal{S}_+$, or $\mathcal{S}_-$.

(8.18) Lemma Let $X \in \mathcal{S}$, where $\mathcal{S}$ is a SSB given as in (8.15). Let $\mathcal{Q}$ be an open set containing X. Then

(a) $I_+(X) \subset I_+[\mathcal{K} - \mathcal{Q}]$ implies $X \in \mathcal{S}_N \cup \mathcal{S}_+$;

(b) $I_-(X) \subset I_-[\mathcal{L} - \mathcal{Q}]$ implies $X \in \mathcal{S}_N \cup \mathcal{S}_-$.

Furthermore, a null geodesic segment on $\mathcal{S}$

(c) passes through X if $X \in \mathcal{S}_N$

(d) has X as future end-point if $X \in \mathcal{S}_+$;

(e) has X as past end-point if $X \in \mathcal{S}_-$.

To prove (8.18), assume, first, that $I_+(X) \subset I_+[\mathcal{K} - \mathcal{Q}]$. Let $\mathcal{B}$ be a small normal coordinate ball with centre X and boundary $\dot{\mathcal{B}}$, where $\mathcal{B} \subset \mathcal{Q}$. Consider a sequence (X_n) of events, with $X_n \in I_+(X) \cap \mathcal{B}$ which converges to X. For each X_n, there exists $W_n \in \mathcal{K} - \mathcal{Q}$ with $W_n \ll X_n$. The timelike curve from W_n to X_n intersects $\dot{\mathcal{B}}$ in Y_n ($Y_n \ll X_n$). Now $\dot{\mathcal{B}}$ is compact, so there is an event Y on $\dot{\mathcal{B}}$ such that a subsequence of (Y_n) converges to Y. The join of Y,X in the normal coordinate system is a geodesic k, through X. If k were spacelike there would be a neighbourhood of X whose past did not intersect the future of some neighbourhood of Y (for small enough $\mathcal{B}$, see (2.8)), contradicting $Y_n \ll X_n$. If k were timelike, we should have $Y \ll X$, so $W_n \ll Y_n \ll X$ for some n, violating $X \notin I_+[\mathcal{K}]$. Thus, k is a null geodesic. Hence $I_+(X) \subset I_+(Y)$. Now, $Y_n \in I_+[\mathcal{K}]$. But if $Y \ll Q$, we have $Y_n \ll Q$ for some n (since $I_-(Q)$ is open), so $I_+(Y) \in I_+[\mathcal{K}]$. Furthermore if $Y \in I_+[\mathcal{K}]$, this would imply $X \in I_+[\mathcal{K}]$ violating (8.15a). Thus $Y \in \mathcal{S}$. This establishes (8.18a), (8.18d) and half of (8.18c). The proof of (18.b), (18.e) and the other half of (8.18c) is exactly similar. (That the two geodesics thus obtained for (8.18c) must coincide, follows since $\mathcal{S}$ is semi-spacelike.)

When two null geodesic segments on $\mathcal{S}$ have an event X in common, this can only happen at $X \in \mathcal{S}_+$ or $X \in \mathcal{S}_-$. This result also holds in the case of "infinitesimally neighbouring" null geodesic segments on $\mathcal{S}$. The following result makes this more precise.

(8.19) Lemma Let $\mathcal{S}$ be a SSB and k a null geodesic segment on $\mathcal{S}$ with a future [resp. past] end-point P. Suppose some open set, containing k-P, meets $\mathcal{S}$ in a smooth null hypersurface for which the convergence ρ is unbounded near P (choosing the tangent vector ℓ^μ to k smoothly at P). Then $P \in \mathcal{S}_+$ [resp. $\mathcal{S}_-$].

The proof of (8.19) that I give depends on another result:

(8.20) Lemma Let k be a null geodesic segment in $\mathcal{M}$ and let $\mathcal{H}_1$, $\mathcal{H}_2$ be smooth null hypersurfaces through k for which $\mathcal{H}_1 \subset I_+[\mathcal{H}_2] \cup \mathcal{H}_2$. Then $\rho_2 - \rho_1 \geqslant |\sigma_2 - \sigma_1|$ on k (ρ_i, σ_i refer to $\mathcal{H}_i$, with the same ℓ^μ for each $\mathcal{H}_i$ at k; i = 1,2).

To prove (8.20), consider two smooth scalar functions u_i (i=1,2) on $\mathcal{M}$ where $u_i = 0$ defines $\mathcal{H}_i$ (in some neighbourhood of k), with u_i increasing towards the future with non-vanishing gradient on $\mathcal{H}_i$. We can scale the u_i's so that $\nabla_\mu u_1 = \nabla_\mu u_2 = \ell_\mu$. In a small enough neighbourhood of k we have $u_2 \geqslant u_1$. Hence, by (5.7), it follows that $x^\mu \nabla_\mu (u_2 - u_1) + 1/2\ x^\mu x^\nu \nabla_\mu \nabla_\nu (u_2 - u_1) \geqslant 0$, on k, for each set of components (x^μ) with $|x^\mu|$ small enough. Let $\nabla_\mu u_i = \ell_{(i)\mu}$, so $\ell_{(1)\mu} = \ell_{(2)\mu} = \ell_\mu$ on k. Choose m_μ, on k, complex, null and orthogonal to ℓ_μ, with $m_\mu \bar{m}^\mu = -1$. Putting $x^\mu = \lambda m^\mu + \bar{\lambda} \bar{m}^\mu$, we get $\lambda^2(\sigma_2 - \sigma_1) + 2\lambda\bar{\lambda}(\rho_2 - \rho_1) + \bar{\lambda}^2(\bar{\sigma}_2 - \bar{\sigma}_1) \geqslant 0$. (cf. (8.6), (8.9)). This must hold independently of λ, and $\rho_2 - \rho_1 \geqslant |\sigma_2 - \sigma_1|$

follows. To prove (8.19), suppose $P \in \mathcal{S}_N$ (with P a future end-point of k). Then, k can be extended into the future on $\mathcal{S}$ to an event $R \in \mathcal{S}$. The past null-cone $C_-(R)$ is non-singular in a neighbourhood $\mathcal{G}$ of k (for R near enough to P). Lemma (8.20) now applies with $\mathcal{N}_2 = \mathcal{G} \cap C_-(R)$ and $\mathcal{N}_1$ as the given null hypersurface. But if $\rho = \rho_1$ is unbounded near P (and positive, since the rays are converging), then $\rho_2 - \rho_1$ must become negative near P, violating (8.20). Thus (8.19) is established.

There are certain further properties of a SSB, $\mathcal{S}$, which can be readily obtained. In particular, $\mathcal{S}$ must be a topological manifold (locally homeomorphic to E^3). To see this, choose $X \in \mathcal{S}$ and set up a smooth timelike vector field in the neighbourhood of X. This vector field can be used to give a homeomorphism of a neighbourhood of X in $\mathcal{S}$ onto a neighbourhood of a smooth spacelike hypersurface near X. This method can also be used to show that any compact portion of $\mathcal{S}$ can be approximated arbitrarily closely by a smooth spacelike hypersurface.

Let us now return to the question of the structure of Cauchy horizons. We noted earlier that the boundary of the set (8.16a) was a SSB (say $\mathcal{S}$) containing $H_+(\mathcal{U})$ as a part. In fact, from (8.18a), it follows that $H_+(\mathcal{U})$ belongs entirely to $\mathcal{S}_N \cup \mathcal{S}_+$. From (8.18 c,d) it then follows that every event on $H_+(\mathcal{U})$ is the future end-point of a null geodesic segment on $\mathcal{S}$. In a similar way, it follows from (8.18b), that the part of $\mathcal{S}$ which bounds $I_-[\mathcal{U}]$ belongs entirely to $\mathcal{S}_N \cup \mathcal{S}_-$ and, from (8.18 c,e), that every event on this part of $\mathcal{S}$ is the past endpoint of a null geodesic segment on $\mathcal{S}$.

We are in a position to establish (8.11) when $\mathcal{H}$ is a closed spacelike hypersurface. For, if such a hypersurface is not a GCH then it has a non-vacuous Cauchy horizon. Without loss of generality, suppose $H_+(\mathcal{H})$ is non-vacuous. Then, as we saw above, every event on $H_+(\mathcal{H})$ is the future end-point of a null geodesic segment on $H_+(\mathcal{H})$, except if the event lies on $\mathcal{H}$. If every maximal

null geodesic meets $\mathcal{H}$, then it follows that $H_+(\mathcal{H})$ must meet $\mathcal{H}$. But this is impossible for a spacelike hypersurface $\mathcal{H}$, since if $P \in \mathcal{H}$ an event just to the future of P will lie in $D_+(\mathcal{H})$.

Note that any $\mathcal{H}$ which is a GCH is also a SSB. This follows because $D_+(\mathcal{H}) = I_+[\mathcal{H}] \smile \mathcal{H}$ here, and $I_+(P) \subset D_+(\mathcal{H})$ implies $P \subset D_+(\mathcal{H})$. Thus, in particular, any GCH is locally homeomorphic to E^3.

As another result, note that if $\mathcal{T}$ is closed and semi-spacelike, the boundary of $I_+[\mathcal{T}]$:

$$(8.21) \qquad \mathcal{S} = \{ X: I_+(X) \subset I_+[\mathcal{T}],\ X \notin I_+[\mathcal{T}] \} ,$$

which is a SSB, consists partly of $\mathcal{T}$, with the property that $\mathcal{S}_- \smile \mathcal{S}_0 \subset \mathcal{T}$. This follows from (8.18). Thus, every event on $\mathcal{S} - \mathcal{T}$ has the property that it is the future end-point of a null geodesic segment on $\mathcal{S} - \mathcal{T}$. Furthermore, $\mathcal{S}$ is locally homeomorphic to E^3.

We now come to the main result. (Here, "future-null complete" means that every null geodesic can be extended into the future to arbitrarily large affine parameter values.)

(8.22) Theorem If a space-time $\mathcal{M}$ satisfies the causality requirements (2.3), (2.4), (2.8), then the conditions

(a) there exists a non-compact GCH, $\mathcal{C}$,

(b) there exists a trapped surface $\mathcal{T} \subset I_+(\mathcal{C})$,

(c) $R_{\mu\nu} \ell^\mu \ell^\nu \leq 0$ for each null vector ℓ^μ,

(d) $\mathcal{M}$ is future-null complete,

are mutually inconsistent.

Proof: Let $\mathcal{S}$ be as in (8.21). Then if $P \in \mathcal{S} - \mathcal{T}$, there will be a null geodesic k, a segment of which lies on $\mathcal{S}$ and has P as a future end-point. Extending the geodesic k maximally into the past, it must eventually intersect $\mathcal{C}$ (since $\mathcal{T} \notin I_-(\mathcal{C})$ and $\mathcal{C}$ is a GCH). Since $\mathcal{S}$ is closed, there must be an event $R \in \mathcal{S}$ which is a past end-point of $k \cap \mathcal{S}$. This must lie on $\mathcal{S}_- \cap \mathcal{S}_0$, so $R \in \mathcal{T}$

CHAPTER 28

Furthermore k must meet the spacelike 2-surface $\mathcal{T}$ orthogonally at R, i.e. lie on one of the two null hypersurfaces which intersect locally in $\mathcal{T}$, since these represent the local boundary of $I_+[\mathcal{T}]$. Thus, $\mathcal{S}$ is generated by the null geodesic segments ($\hat{k}$) which are orthogonal to $\mathcal{T}$ at a past end-point, and which may have a future end-point (where they encounter $\mathcal{S}_+$). In fact, any such segment $\hat{k}$ must have a future end-point. For, if we choose the tangent vector ℓ^μ to be parallelly propagated along $\hat{k}$, we have, by (6.28a), (8.22a) and by (8. 7): $D(a^{1/2}) = - a^{1/2} \rho$,

$$(8.23) \qquad D^2(a^{1/2}) = D(-\rho a^{1/2}) = a^{1/2}(1/2\, R_{\mu\nu} \ell^\mu \ell^\nu - |\sigma|^2) \leqslant 0.$$

But $a^{1/2}$ is initially decreasing (by (8. 7)), so, by (8.23) it must decrease to zero.[97] This happens at a finite value of an affine parameter on k, since D is differentiation with respect to such a parameter. (When $a^{1/2}$ becomes zero, k has encountered a caustic, i.e. k intersects another "neighbouring" k.) Since $\rho = - 1/2\, D (\log a)$, ρ must become unbounded at this event, Q ($Q \in \mathcal{M}$ by (8.22d)). Thus, it follows from (8.19) that k has a future end-point. (In fact this end-point generally occurs before the caustic is reached, namely at a "crossing region" of $\mathcal{S}$.) Now Q varies continuously as k varies, so the finite segments RQ of k generate a compact set. Thus $\mathcal{S}$, being a closed subset of a compact set, is also compact. So, we have that:

(8.24) $\mathcal{S}$ is a semi-spacelike compact topological 3-manifold in $\mathcal{M}$.

Now, it is a theorem[98] that any space-time admits a smooth timelike vector field. We can use the trajectories of this vector field to achieve a one-to-one mapping of $\mathcal{S}$ into $\mathcal{C}$ (both sets are semi-spacelike). This is impossible since the dimensionalities are the same but $\mathcal{S}$ is compact and $\mathcal{C}$ is non-compact. Thus (8.22 a, b, c, d) are mutually inconsistent.

We may ask which of the conditions of (8.22) is most likely to break down. Condition (8.22c) comes from the requirement that if the energy-momentum tensor is given by Einstein's equations (7.9) (with or without cosmological term), we should like to be able to assert that the energy density is necessarily non-negative:

$$(8.25) \qquad E = T_{\mu\nu}\, t^{\mu} t^{\nu} (t_{\alpha}\, t^{\alpha})^{-1} \geqslant 0 \ ;$$

for any observer (t^{μ} being tangent to his world-line). Condition (8.25) would only be violated by very "unreasonable" matter. (In an "eigen-tetrad" of $T_{\mu\nu}$, (8.25) is equivalent to $T_{00} \geqslant 0$, $T_{00} + T_{11} \geqslant 0$, $T_{00} + T_{22} \geqslant 0$, $T_{00} + T_{33} \geqslant 0$.) We obtain (8.22c) as a limiting case of (8.25) by allowing $t^{\mu} \longrightarrow \ell^{\mu}$. It would be interesting to know if a violation of (8.25) is in any way "forced" upon us by considerations such as (8.22), since such cosmological theories as the Steady-State or Hoyle-Narlikar theories (8.25) violation is required.[73,99] However, there seems to be a serious problem in connection with quantum field theory, if (8.25) is violated, namely an apparent catastrophic instability of the vacuum. But the issue does not seem to be completely clear-cut, to my mind.

The condition (8.22d) of future-null completeness is not an unreasonable one. (It is somewhat weaker than geodesic completeness.) In crude terms it states that "photons do not just disappear". Without some kind of completeness or other global condition, no inference concerning the existence of singularities can ever be made. For, if any closed set is removed from $\mathcal{M}$ (which leaves it connected) we still have a space-time manifold. It is just that completeness would be destroyed. However, one can, perhaps, question whether null completeness (or geodesic completeness) is the really satisfactory physical condition to impose. I have in mind the model of the steady-state universe. This being the future half of a de Sitter hyperboloid, it is not past-null complete (cf.

section VII). This is not a feature that appears to have worried cosmologists particularly, and the steady-state model is one of those which has been most seriously considered as a model of the actual universe.[73] The time-reverse $\mathcal{W}$ of the steady-state model would thus violate the condition (8.22d); but would we have to rule out such a model on these grounds? Actually there are two complicating features here. A photon in $\mathcal{W}$ would not actually "just disappear" at the final boundary; in fact, it encounters an infinite matter density at this boundary. The reason this infinite matter density does not show up as a singularity in the space-time is that its effect has to be compensated by an infinite negative energy density of the creation field. It appears that this type of effect could not occur with a strict adherence to (8.25). In addition, if we consider all the details of the irregularities of matter, it is by no means clear that a "realistic" steady-state model would be past-null incomplete. A null geodesic into the past would be deflected infinitely many times before reaching the boundary. Should this be counted as a singularity in the past? The time-reverse of this type of behaviour has relevance to the collapse problem as we shall see shortly. Possibly of a somewhat similar nature is the type of incompleteness encountered with the Taub-NUT– Misner universe.[100] Here we have incompleteness and inextendibility but (in a well-defined sense) no singularity. All models which exhibit this type of (Misner) incompleteness must violate (2.8), however.[101]

The possibility of a failure of (8.22a) is something which now must be considered. There are two ways that (8.22a) can fail to be satisfied. In the first instance, the universe might be "closed". Then, according to the usual ideas, this would mean the existence of a compact GCH. With this condition in place of (8.22a), the proof of (8.22) would break down at the final stage. But it seems hard to believe that the question of whether the universe, as a whole,

is open or closed, should seriously affect the behaviour of a "local" collapsing system (which would be smaller by a factor of at least 10^{-15}). Indeed, an examination of the proof of (8.22) shows that the argument can still be made to carry through in the case of a compact GCH unless, in a well-defined sense, the collapsing object eventually "swallows up" the entire universe! However, a closed universe (like the Friedmann model) which eventually recontracts to a highly condensed state could, in principle, satisfy this condition. But the elimination of singularities by this means seems to be ruled out by a theorem of Hawking, which states, in effect, that "almost all" space-times with a compact GCH must encounter singularities provided a slightly stronger (but still "reasonable") condition on $T_{\mu\nu}$, than (8.22c), holds, namely

$$(8.26) \qquad T_{\mu\nu}\, t^{\mu} t^{\nu} \geqslant 1/2\, T^{\mu}{}_{\mu}\, t^{\nu} t_{\nu}$$

for each timelike vector t^{μ}. (Here $T_{\mu\nu}$ is defined without cosmological constant λ, cf. (7.9).)[78]

The second way in which (8.22a) might be violated is, to my mind, the more serious possibility. The "Laplacian" idea that the future of the universe should be completely determined by its behaviour at one "time" predates both relativity and quantum theory. We have become accustomed to non-determinism in the latter theory, so why should we still demand determinism in all cases for general relativity? The possibility that the universe should not possess a GCH at all is an intriguing one. There are, in fact, certain physically interesting space-times which do not possess GCH's. Notable examples are the gravitational or gravitational-electromagnetic plane wave[30a] space-times. Other examples which are more relevant to the question of collapse (although, unlike the plane waves they possess singularities) are the Kerr[91] and Reissner-Nordström[102] space-times, as extended, respectively, by Boyer-Lindquist[88] and Graves-Brill.[103] The Kerr metric is:

$$(8.27a)\quad ds^2 = dv^2 - \frac{2mr}{h}(dv + a\sin^2\theta\, d\varphi)^2 - 2\,dv\,dr + h\,d\theta^2 + 2a\sin^2\theta\, dr\, d\varphi + (r^2 + a^2)\sin^2\theta\, d\varphi^2$$

where m and a are constant and where

$$(8.27b)\quad h = r^2 + a^2\cos^2\theta\ .$$

The Reissner-Nordström metric (with advanced time v and electric charge defined by e) is

$$(8.28)\quad ds^2 = \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right) dv^2 - 2\,dv\,dr - r^2(d\theta^2 + \sin^2\theta\, d\varphi^2).$$

The metric (8.27) with $m > |a|$ and the metric (8.28) with $m > |e|$ are somewhat similar in that they possess, not only an event horizon (Schwarzschild "throat"), given, respectively by

$$(8.29a)\quad r_+ = m + (m^2 - a^2)^{1/2},\quad r_+ = m + (m^2 - e^2)^{1/2},$$

but also a second "horizon" at which the null cones tip back again:

$$(8.29b)\quad r_- = m - (m^2 - a^2)^{1/2},\quad r_- = m - (m^2 - e^2)^{1/2}.$$

Both metrics have true singularities at $r = 0$, but these differ somewhat in structure. In the case of the Kerr metric the singularity has a ring structure. By passing through the ring one enters a region with $r < 0$. Ignoring the nature of the metrics near (and through) the singularities, the situation is as in fig. 19. By patchwork similar to that in the Schwarzschild case, but here more complicated, a "maximal" analytic extension is obtained: fig. 20. The shaded portion in fig. 20 represents the whole of fig. 19.

Now, the significance of all this is that by following the star inwards and then (if necessary) giving himself an acceleration outwards after crossing $r = r_+$, an observer can find himself in a portion of the universe not causally

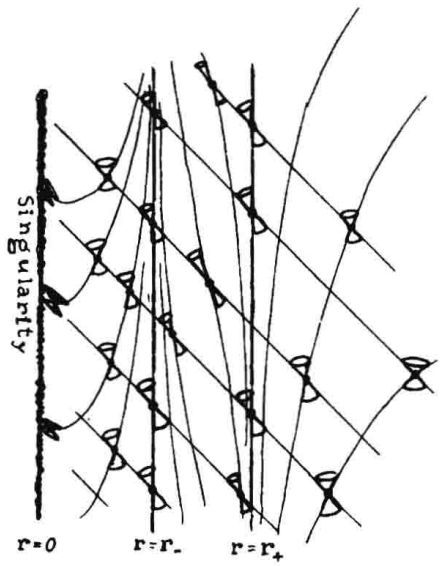


fig. 19: Kerr picture (or Reissner-Nordstrom)

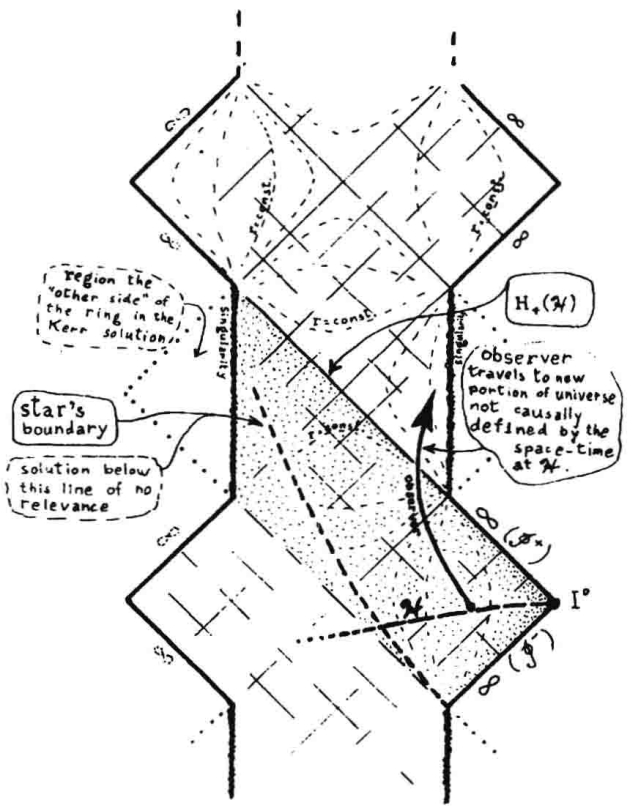

region the "other side" of the ring in the Kerr solution
star's boundary
solution below this line of no relevance
r=const.
observer travels to new portion of universe not causally defined by the space-time at 𝒲.
I°

fig. 20: Carter-Boyer-Lindquist picture (incorporating Graves-Brill)

determined by information on an apparently reasonable initial hypersurface $\mathcal{H}$. In fact he crosses $H_+(\mathcal{H})$. The region he enters actually contains a singularity which he can "see". But of course, since he has crossed $H_+(\mathcal{H})$, we cannot guarantee that the space-time he enters is in fact the one that we have constructed for him. In effect, "another universe" has joined on to the universe he started from. The particular continuation we have chosen was obtained here by an analytic continuation rather than by the use of field equations. (But even analytic extensions do not ensure uniqueness globally. A striking example occurs with the Taub-NUT-Misner space-time, but also in Minkowski space-time analytic extension does not imply uniqueness. The difficulty is that one does not know, in general, when to identify events reached in different ways.) Even if we can avoid the singularity in the "new universe" there is a problem of principle here, because, like in the early stages of a "big bang" cosmology, causally disconnected regions have to come together in such a way that they "fit". A striking example of this occurs with the Reissner-Nordström solution, since the "new universe" has to possess charged particles of just the right total charge. (This follows even independently of any symmetry assumptions.)

There is a further difficulty confronting our observer who tries to cross $H_+(\mathcal{H})$. As he looks out at the universe that he is "leaving behind", he sees, in one final flash, as he crosses $H_+(\mathcal{H})$, the entire later history of the rest of his "old universe". It is here, in fact, that cosmological questions might be very relevant to him. If, for example, an unlimited amount of matter eventually falls into the star then presumably he will be confronted with an infinite density of matter along $H_+(\mathcal{H})$ (if $H_+(\mathcal{H})$ still exists), like in the case of the time-reversed steady-state model. If a large, but finite amount of matter falls into the star is it clear that the observer can cross $H_+(\mathcal{H})$?

And if non-singular regions like $H_+(\mathscr{H})$ really can be crossed, is there any reason to believe that this will enable us to avoid having to have true singularities in the space-time? A result of Hawking suggests that it will not: Some fully general space-times (subject to (8.26)) satisfying only an inequality (converging normals) on a compact spacelike hypersurface, necessarily develop singularities irrespective of any causality conditions whatsoever.[101]

In many ways it is appealing to think, the presence of singularities notwithstanding, that our universe may not contain any GCH. For if there is a GCH, $\mathscr{H}$, then the topology of the universe is comparatively uninteresting, being $\mathscr{H} \times E^1$ (as is not hard to show). A promising place for developments in the topology to occur, would seem to be in gravitational collapse. We saw in the proof of (8.22) that from a region with an initially uninteresting topology ($\mathscr{C} \cong E^3$), might develop a "sub-universe" with a compact spatial topology ($\cong \mathscr{S}$; see (8.24)). Thus, in effect we might envisage a "new universe" branching off the old one. If such a system could be self-sustaining, we might picture a kind of self-generating universe: After an initially compressed stage in which new baryons are created, each branch would undergo rapid expansion. Eventually, collapsing portions would again branch off. The entire universe would have a kind of tree structure, but observers would be directly aware only of the branch on which they resided. They could, however, travel from one branch to a "sub-branch" by dropping through a Schwarzschild "throat" and accelerating outwards. By allowing different portions of the branches to be contracting and expanding simultaneously, we could, if desired, evade all the presently known theorems predictng singularities.

Such a model perhaps seems artificial. It is presented here mainly as an example to indicate that it may not be possible to represent even the large-scale structure of the universe by means of simple smoothed out models. The detailed implications of Einstein's general theory -- or of some extended theory if this turns out to be required -- would then be of a very real relevance to cosmology.

Appendix: Combinatorial Quantum Theory and Quantized Directions

According to conventional quantum theory, angular momentum can take only integral values (measured in units of $1/2\ \hbar$) and the (probabilistic) rules for combining angular momenta are of a combinatorial nature. Also, according to quantum theory, a system with zero total angular momentum must be spherically symmetrical. A system must, thus, involve a relatively large angular momentum in order to determine a well-defined direction in space, and we may picture the axis of the angular momentum as giving such a direction. In the limit of large angular momenta, we may therefore expect that the quantum rules for angular momentum will determine the geometry of directions in space. We may imagine these directions to be determined by a number of spinning bodies. The angles between their axes can then be defined in terms of the probabilities that their total angular momenta (i.e. their "spins") will be increased or decreased when, say, an electron is thrown from one body to another. In this way the geometry of directions may be built up, and the problem is then to see whether the geometry so obtained agrees with what we know of the geometry of space and time.

Although there is a standard procedure for the treatment of such a problem (in the non-relativistic case: namely the use of 3-j and 6-j symbols, etc.), it is convenient to make use of an alternative (but equivalent) formalism which can be described very briefly as follows. Consider, first, an n-dimensional Kronecker delta δ_{ab}. Then a set of "isotropic" Cartesian tensors can be built up from this one symbol and scalars by means of the operations of addition, outer multiplication, transposition of indices, and contraction. The only identities satisfied by δ_{ab} which hold independently of n are $\delta_{ab} = \delta_{ba}$, $\delta_{ab}\delta_{bc} = \delta_{ac}$ (summation convention assumed). Also, we have $\delta_{aa} = n\ (> 0)$ and a certain identity, which depends on n, holds. Now consider an abstract structure which behaves formally exactly like the above set of tensors except that the formal equation $\delta_{aa} = -2$ is imposed, together with the one identity: $\delta_{ab}\delta_{cd} + \delta_{ac}\delta_{db} + \delta_{ad}\delta_{bc} = 0$. The "scalars" of this abstract structure will be taken to be the rational numbers. The tensor-like quantities which can be constructed in this way I call binors. It can then be shown that a necessary and sufficient condition for a binor $A_{ab\ldots d}$ to vanish is that its norm $\|A \ldots\| = |A_{ab..d}A_{ab..d}|$ (a rational number) should vanish.

To obtain the physical interpretation of the binors, I envisage the following type of situation. Imagine the universe to be represented as a network of segments

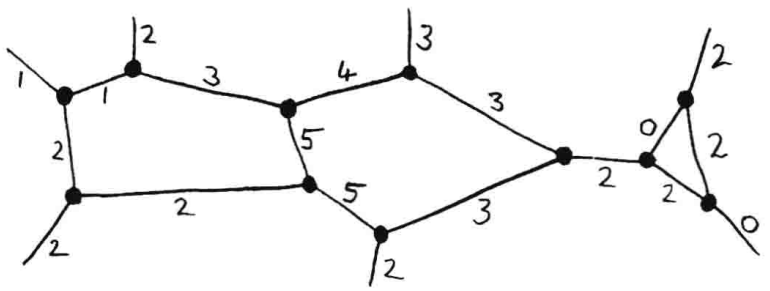


fig. A1

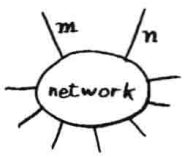
(see fig. A1) where, for simplicity, it will be assumed that each segment connects two vertices and each vertex is a junction of just three segment ends. Associated with each segment is a non-negative integer. Each segment is to be thought of as representing the world line of a particle, nucleus, atom, etc. Or, generally, some structure which may be momentarily considered as stationary and isolated from the rest of the universe. The network thus gives a kind of combinatorial space-time picture of the universe. The integer associated with each segment is twice the total intrinsic angular momentum quantum number of the particle (or structure). Hence, for a pi-meson or ground state helium atom this integer is zero; for an electron or proton it is one; for a deuteron it is two; and so on. In order for the proposed calculus to give an accurate description of a part of the universe, it must be supposed that it is possible to neglect effects due to relative velocities of the particles (or structures). In this sense the theory is a non-relativistic one.

Corresponding to each possible such (open) network is associated a binor which is a contracted product of binors described as follows. Each segment numbered 0 is represented by the scalar 1. Each segment numbered 1 is represented by a δ_{ab}. Segments marked 2 by $\delta_{ac,\,bd} = 1/2!\,(\delta_{ab}\delta_{cd} - \delta_{ad}\delta_{cb})$, segments marked 3 by

$$\delta_{abc,\,def} = 1/3! \begin{vmatrix} \delta_{ad} & \delta_{ae} & \delta_{af} \\ \delta_{bd} & \delta_{be} & \delta_{bf} \\ \delta_{cd} & \delta_{ce} & \delta_{cf} \end{vmatrix}$$

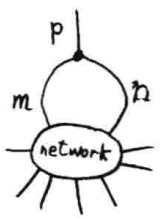
and so on. The first group of indices of $\delta_{ab..d,ef..h}$ is to be associated with one end of the segment and the second group with the other end. (This is symmetrical since $\delta_{ab} = \delta_{ba}$ implies $\delta_{ab..d,ef..h} = \delta_{ef..h,ab..d}$.) At each vertex the three relevant groups of indices involved must all be paired off and contracted so that none of these indices remains uncontracted, and no two belonging to the same $\delta_{...,...}$ are contracted together. This implies that the sum of the three integers involved is even, and is at least twice the greatest of them. Then, the result of these contractions will be unique up to sign. The free indices of the resultant binor are then just those corresponding to the free ends of the network.

Consider, now, a situation in which a portion of the universe is represented by a known network (fig. A2) having, among its free ends, one numbered m and one numbered n. Suppose the particles or structures represented by these two ends combine together to form a new structure. (See fig. A3) We wish to know, for



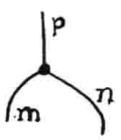
A....

fig. A2.



B....

fig. A3.



C....

fig. A4.

D....

fig. A5.

any allowable p, what is the probability that the angular momentum number of this new structure be p. Fig. A4. is the network representing the combining of these two structures to form the third, but ignoring the rest of the universe; fig. A5 is the "network" for the final structure alone. Let A..., B..., C..., D..., be the binors representing the networks of figs. A2, A3, A4 and A5, respectively. Then,

$$\underline{\text{required probability}} = \frac{\| B_{...} \|}{\| A_{...} \|} \times \frac{\| D_{...} \|}{\| C_{...} \|}$$

(in fact $\| D_{...} \| = (-1)^p (p+1)$ and $\| C_{...} \| = (-1)^{(m+n+p)/2} \left(\frac{m+n-p}{2}\right)! \left(\frac{n+p-m}{2}\right)! \left(\frac{p+m-n}{2}\right)! \left(\frac{m+n+p+2}{2}\right)! / m!n!p!$). As a particular case of this we can deduce the result that the binor corresponding to a network vanishes if and only if the situation is "forbidden" according to the rules of quantum theory.

Consider now the situation of fig. A6 involving two bodies with large angular momenta M and N. We might define the angle between the axes of the bodies in terms of the relative probability (defined in terms of an ensemble

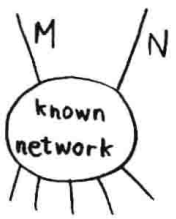


fig. A6.

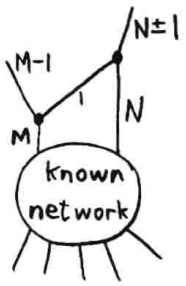


fig. A7.

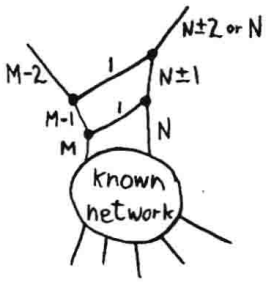


fig. A8.

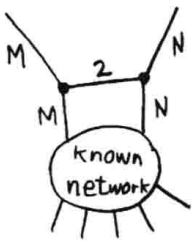


fig. A9.

of systems with the same fig. A6 network) of occurrence of N+ 1 and N -1, respectively, in the "experiment" given by fig. A7. However, part of this probability may be due to ignorance of the relationship between the bodies. (This may manifest itself in the absence of sufficient connecting links in the know network.) To eliminate the possibility that part of the probability be due to ignorance we envisage a repetition of the "experiment" as given in fig. A8. If the probability given in the second experiment is essentially unaffected by the result of the first experiment then we may say that the angle θ between the axes of the bodies is well-defined and is determined by this probability. It then turns out that the binor of fig. A9 is essentially $1/2 \cos \theta$ times the binor of fig. A6 and from this fact, and certain binor identities, it is possible to show that the angles obtained in this way satisfy the same laws as do angles in a three-dimensional Euclidean space.

This is very satisfactory and is perhaps a little surprising in at least two respects. In the first instance, since no complex numbers were used in the binor calculus it is somewhat remarkable that a full array of directions in three-dimensional space has been built up, rather than in, say, just a two-dimensional subspace, since in ordinary quantum theory, in order to build up all the wave functions for all the possible spin directions for an electron, complex linear combinations must be used. Secondly, according to standard quantum theory, the wave function of a system of high angular momentum will not normally determine a well-defined axis in space contrary to what has apparently been assumed here.

The answer to both these points seems to lie in the fact that the "directions" that emerge in the theory described here are things which are defined by the systems in relation to one another and they will not generally agree with the directions in a previously given (and unnecessary!) background space. The space that is obtained here is to be thought of (indeed must be thought of) as being the one determined by the systems themselves.

It is to be hoped that some modification to the above scheme might enable the effects due to relative velocities of systems to be taken into account so that perhaps a four-dimensional space-time might be constructed. (Time is absent from the above theory even to the extent that the time ordering of events is irrelevant!) Two additional features would have to be involved. The first is that the relativistic addition of angular momenta includes the possibility of multiple pair creation and many of the additional complications implied by relativistic field theory. The second is that relative velocity implies the possibility of a mixing of spin with orbital angular momentum so that the idea of "distance" between the world lines of particles is involved. Particularly because of this second feature, the fully relativistic theory would seem to be of a different order of difficulty than the one treated above.

REFERENCES and FOOTNOTES

1. B. Riemann, 1854, "Uber die Hypothesen welche der Geometrie zu Grunde liegen," in Gesammelte Mathematische Werke (H. Weber, ed.), Dover Publications, New York, 1953.

2. This idea was made somewhat more explicit by W.K. Clifford, Proc. Camb. Phil. Soc., 2, 157 (1876). For a significant modern viewpoint, see J.A. Wheeler, Geometrodynamics, Academic Press, New York, 1962.

3. A. Einstein, Ann. Physik. 94, 769 (1916).

4. See, for example, S. Helgason: Differential Geometry and Symmetric Spaces, Academic Press 1962, New York.

5. It is not intended to be suggested that Newton actually viewed space-time in this way. From some of his writings, one might even infer that his views would incline, perhaps, more to the Aristotelian, rather than to the Galilean view of space-time. However, it seems to me to be, at least, possible that his private views were more "relativistic" than one might infer from these writings. His dogged persistence in hanging on to a particle theory of light, in the face even of his own discoveries of some of its wave properties -- leading to an almost uncanny foresight of a wave-particle duality -- must surely have had some very strong physico-intuitive motivation. Such a motivation could have been supplied by a desire for Galilean(?) invariance coupled with the observation that an all-pervading ether would destroy such invariance.

6. E. Cartan, Ann. Ec. Norm. Sup., 40, 325 (1923); 41, 1 (1924); K. Friedrichs, Math. Ann., 98, 566 1927); C. Kilmister, J. Math. Mec., 12, 1 (1963); A. Trautman, Comptes Rendus, 257, 617 (1963); in Perspectives in Geometry and Relativity, Indiana University Press, Bloomington and London (1966); in Lectures on General Relativity (1964 Brandeis Summer Institute in Theoretical Physics, Vol. I), Prentise-Hall, Englewood Cliffs, 1965.

6a. Effectively, it is the "weak" equivalence principle which is being appealed to here.

7. Strictly, perhaps, one should say "proportional", but the constant of proportionality is simply absorbed in the definition of the gravitational constant. The most accurate experimental verification, due to P.G. Roll,

R. Krotkov and R.H. Dicke, shows that gold and aluminium accelerate equally towards the sun to an accuracy of a part in 10^{11}. Eötvös, in 1890 and later, obtained an accuracy of a few parts in 10^{9}. For details see R.H. Dicke, in Relativity Groups and Topology (Les Houches 1963) Gordon and Breach, New York 1964.

7a. It is sometimes argued that the name "relativity" is somewhat inappropriate for Einstein's theory, since it deals with an "absolute" space-time. While this is perhaps true, one could reasonably say that any theory becomes "absolute" once it is well enough formulated. The point about Einstein's theories was that certain concepts, notably "time", while previously regarded as "absolute", become "relative" in the new theory.

8. Since I wrote this, a recent experimental result due to Dicke and his collaborators has been announced, which seems to tell against the accuracy of general relativity. According to this result, the sun's quadrupole moment is measured as being large enough to account for 7% of the observed perihelion advance of Mercury, thus destroying the accuracy of the general relativistic prediction by this amount. If the result is borne out by future analysis, then general relativity could not be maintained precisely in its present form. The modification suggested by Dicke would not, to my mind destroy the validity of the "Einsteinian view", but merely modify it. The basic geometry of space-time would then be, more properly, a conformal structure than a Riemannian one (cf. the remarks at the end of section II).

9. A. Einstein, The Meaning of Relativity, 5th ed., pp. 133-166. Princeton University Press, Princeton, 1955. E. Schrödinger, Space-time Structure, Cambridge University Press, Cambridge, 1950. V. Hlavatý, Geometry of Einstein's Unified Field Theory, P. Noordhoff, Gröningen, 1958.

9a. Kaluza, S. Preuss. Akad. Wiss., 966 (1921); O. Klein, Z. Physik, 73, 895 (1926).

10. D.W. Sciama, in Recent Developments in General Relativity, Pergamon Press and PWN, Warszawa, 1962; T.W.B. Kibble, J. Math. Phys. 2, 212 (1961).

11. P. Jordan, Ann. Physik, 219 (1947); Y. Thiry, "Etude mathématique des équations d'une théorie unitaire à quinze variables de champ" (thesis), Gauthiers-Villars (1951); R.H. Dicke, Phys. Rev. 125, 2163 (1962); C. Brans and R, H. Dicke, Phys. Rev. 124, 3(1961).

12. Much has been written on such "Lorentz-covariant" formulations of general relativity (cf. N. Rosen, 1940, Phys. Rev. 57, 147; S. Gupta, 1957, Rev. Mod. Phys. 29, 334; W.E. Thirring, 1961, Annals of Phys., 16, 96-117). There is, for some purposes, a certain advantage in using such a formulation, for example in direct approaches to the problem of quantization, since one can appeal to pre-existing knowledge concerning Lorentz- (i.e. Poincaré-) covariant field theories. It is often argued that such a formulation is equivalent to the curved space-time approach. However, this is not correct if Lorentz causality is to be maintained, which would require that signals cannot propagate faster than light in the background space-time. For example, in the case of a gravitational plane wave, it is impossible to arrange this (cf. ref. 30a). (of course, non-Euclidean topologies also cannot be treated using such a formulation.)

13. L.E.J. Brower (see A. Heyting, Intuitionism, North-Holl. Amsterdam, 1956). H. Weyl, Das Kontinuum, Chelsea, New York, 1956; A.M. Turing, Proc. Lond. Math. Soc. (2) 42, 230 (1936-37); H.G. Rice, Proc. Amer. Math. Soc. (2) 52, 784 (1954).

13a. A simple example of an increasing sequence of computable numbers, bounded above, whose limit is non-computable, is

$$a_r = \sum_{n=1}^{\infty} 2^{-n} f_r(n) ,$$

where $f_r(n)$ is 1 or 0 according as the n^{th} Turing machine has or has not stopped by the r^{th} step.

14. E. Schrödinger, Science and Humanism, Cambridge University Press, Camb., 1956.

15. An alternative point of view, that non-constructiveness in a somewhat different context might even be desirable, has been put forward by A. Komar, Phys. Rev. (2) 133, B542 (1964).

16. See O. Klein, Nuov. Cim. Suppl., 6, 344 (1957); L. Landau in Niels Bohr and the Development of Physics, Pergammon, London, 1955; S. Deser, Revs. Mod. Phys. 29, 417, (1957); B.S. De Witt, Phys. Rev. Lett. 13, 114 (1964). However, in its present state, the quantum theory of gravitation is very far from removing the divergences which appear in the theory itself, let alone those of any other theory.

17. But, as DeWitt has emphasized, quantum gravitational effects would not be expected to become important until densities (or the corresponding curvatures) were reached, of the order of that of the "entire universe" (10^{40} protons) concentrated in a sphere of the radius of one proton.

18. This is ignoring the possibility of a superselection rule preventing this (G.C. Wick, E.P. Wigner and A.S. Wightman, Phys. Rev., 88, 101 (1952)). For an interesting argument against superselection rules in general, see Y. Aharonov and L. Susskind, "Charge Superselection Rule", to appear.

19. For an account of this, see the work by Wheeler referred to in Footnote 2; also C.W. Misner and J.A. Wheeler, Annals of Physics 2, 525 (1957).

20. A. Schild, Phys. Rev. 73, 414 (1948).

21. Y. Ahmavaara, J. Math. Phys. 6, 87 (1965).

21a. By replacing Schild's integral lattice by a lattice based on the rational field, E.L. Hill (Phys. Rev. 100, 1780 (1955)) was able largely to remove this particular problem. However his solution seems to be rather artificial as it loses the advantage of having a minimum length.

21b. H.S. Snyder, Phys. Rev. 71, 38 (1947).

22. G.F. Chew, Sci. Prog. 51, Oct. 1963.

23. R. Penrose, "Twistor Algebra" (to appear in J. Math. Phys. 1967).

Footnotes 24-28 do not exist

29. E.H. Kronheimer and R. Penrose, "On the Structure of Causal Spaces" (to appear in Proc. Camb. Phil. Soc., 1967).

30. It is, in fact, a theorem (L. Marcus, Ann. Math. 62, 411 (1955)) that any hyperbolic normal manifold not possessing this property has a double covering manifold for which this property does hold.

30a. R. Penrose, Revs. Mod. Phys. 37, 215 (1965).

31. Some approaches to unified field theory would seem to amount to little more than this!

32. Of course, if we use Minkowskian coordinates in flat space-time, the components of the metric tensor are just ±1 so that we need not notice that the g field enters. But we cannot avoid the explicit appearance of g when we use a generally covariant formalism, even in flat space-time.

33. One possibility is to define g simultaneously with introducing spinors (P.G. Bergmann, Phys. Rev. 107, 624 (1956)). The point, here, is that we cannot have spinors without having a metric (or, at least, a metric up to proportionality).

34. It is, however, possible to describe a spinor field in tensorial terms (E.T. Whittaker, Proc. Roy. Soc. (Lond.) 158A, 38 (1947)); H.S. Ruse, Proc. Roy. Soc. (Edinb.) A62, 64 (1944), but in a way which is more complicated and which destroys the linear nature of the spinor wave equations, not to mention the two-valuedness of the spinors. Strictly speaking, though, the differing role of the g field in relation to spinor fields, or to complicated tensor fields, is not really quite so clear-cut as I have suggested in the text.

35. R. Brauer and H. Weyl, Am. J. Math. 57, 425 (1935); E. Cartan, Leçons sur la Théorie des Spineurs (Actualites Scientifiques et Industrielles 643, 701) Hermann, Paris, 1938.

36. The fact that space- and time-reflections are not represented directly by (2.18) is, to my mind, if anything an advantage, now that it is known that space-reflection invariance and very possibly also time-reflection invariance are not exact symmetries of nature.

37. See ref. to Whittaker in ref. 34; also, W.T. Payne, Am. J. Phys. 20, 253, (1952).

38. J. W. Milnor, Enseignement Math. (2) 9, 198 (1963).

39. Y. Aharonov and L. Susskind, "Rotations by 2π" (to appear).

40. B.L. van der Waerden, Nachr. Akad. Wiss. Gottengen, Math.-physik. Kl. 100 (1929).

41. R. Penrose, Ann. of Phys. 10, 171 (1960).

42. In the original treatment of L. Infeld and B.L. van der Waerden (S.B. Preuss. Akad. 9, 380 (1933)) the charge of a spinor is defined in terms of the nature and positions of its indices. Thus the spin and the charge have to be related to one another, so that for example it becomes impossible to treat a neutron or a charged pion in terms of the general formalism. The present treatment differs from theirs in that the charge concept is introduced completely independently of the spinor index structure.

43. O. Laporte and G.E. Uhlenbeck, Phys. Rev. 37, 1380 (1931).

44. The concept of an exact set was introduced by the author in a widely circulated preprint entitled "Null Hypersurface initial data for Classical Fields of Arbitrary Spin and for General Relativity". This article can be found in P.G. Bergmann's A.R.L. Tech. Documentary Rep. 63-56 (1963). Sections IV and V are based on an extended version of part of this article.

45. See ref. 43; also, W.L. Bade and H. Jehle, Revs. Mod. Phys. 25, 714 (1953).

46. R. Penrose, Proc. Roy. Soc. A284, 159 (1965).

47. The general method of obtaining the unsymmetrized derivatives in terms of symmetrized derivatives follows that of ref. 41, to which the reader is referred for details.

48. In general relativity proper, there are further difficulties because (except for gravitation itself) a local energy-momentum tensor is required. A zero rest-mass field of spin $s \geqslant 3/2$ does not possess such a local energy-momentum tensor. In the case of gravitation itself this problem does not arise since the gravitational energy does not appear in the energy-momentum tensor but emerges naturally as a non-local quantity. In ref. 46, the curved space-time analysis was based on (4.5), which enables gravitation to be treated, in addition to zero rest-mass fields of spin $s \leqslant 1$. In connection with eq.(4.17), see H.A. Buchdahl, Nuovo Cim. 10, 96 (1958); 25, 486 (1962) and J. Plebanski, Acta Phys. Polon. 27, 361 (1965).

49. V. Fock, Z. Sowjetunion 12, 404 (1937), R.P. Feynman and M. Gell-Mann, Phys. Rev. (2) 109, 193 (1958).

50. P.A.M. Dirac, Proc. Roy. Soc. A155, 447 (1936), M. Fierz, Helv. Phys. Acta, 12, 3 (1939).

51. M. Fierz and W. Pauli, Proc. Roy. Soc. (Lond.) A173, 211 (1939).

52. See the articles of Buchdahl given in ref. 48.

53. There may, however, be additional global topological considerations involved, especially with general relativity, since then the structure of the space-time itself is also being determined. There is the question of when events arrived at by different routes should be considered as identical.

54. Compare also J.L. Synge, Relativity: The General Theory, North-Holland, Amsterdam, 1960.

55. See, for example, J. Hadamard, Lectures on Cauchy's Problem in Linear Partial Differential Equations, Dover, New York, 1952; G.F.D. Duff, Partial Differential Equations.

56. See E.M. Corson, Introduction to Tensors, Spinors, and Relativistic Wave Equations, Blackie, Glasgow, 1953; p. 121.

57. See the articles by Y. Bruhat and by R. Arnowitt, S. Deser and C.W. Misner in Gravitation: An Introduction to Current Research, ed. L. Witten, Wiley, New York, 1962.

58. See F.A.E. Pirani and A. Schild, Boll. Acad. Pol. Ser. Sci. 9, 543 (1961); also, article in Perspectives in Geometry and Relativity, Ed. Hoffmann, Indiana University Press, Bloomington; K.L. Stellmacher, Math Annalen, 123, 34 (1951); R. Penrose, in Perspectives in Geometry and Relativity (above).

59. Spin Coefficients were first introduced by E.T. Newman and R. Penrose, J. Math.Phys. 3, 566 (1962); 4, 998 (1963).

60. These equations are due to appear in an article by A. Exton, E.T. Newman, and R. Penrose, submitted to J. Math. Phys. They are developments of equations taken from E.T. Newman and T.W.J. Unti, J. Math. Phys. 3, 891 (1962) and from B. Kozarzewski, Acta Phys. Polon. 27, 775 (1965), which are in turn developments of equations given in ref. 59.

61. H. Bondi, Nature, Lond. 186, 535 (1960), H. Bondi, M.G.J. van den Burg and A.W.K. Metzner, Proc. Roy. Soc. A269, 21 (1962), R.K. Sachs, Proc. Roy. Soc. A270, 103 (1962); compare also Newman-Unti (ref. 60); see also F.H.J. Cornish, Proc. Roy. Soc. A282, 372 (1964), R. Penrose, Phys. Rev. Lett. 10, 66 (1963) and ref. 66.

62. E.T. Newman and R. Penrose, Phys. Rev. Lett. 15, 231 (1965) and another article to be submitted to Proc. Roy. Soc. (Lond.); also, Exton-Newman-Penrose (ref. 60).

63. E.T. Newman and R. Penrose, J. Math. Phys. 7, 863, (1966).

64. H.S. Ruse, Proc. Roy. Soc. Edinb. 57, 97 (1937).

65. See, for example, A. Janis and E.T. Newman, J. Math. Phys. 6, 902 (1965).

66. R. Penrose, in Relativity, Groups and Topology, Ed. C. DeWitt and B. DeWitt, Gordon and Breach, New York, 1964.

67. R.K. Sachs, Phys. Rev. 128, 2851 (1962).

68. See, for example, N.H. Kuiper, Ann. of Math. 50, 916 (1949); H. Rudberg, "The Compactification of a Lorentz Space" (thesis) Uppsala, 1957; R. Penrose, in Proceedings on Theory of Gravitation, Conference in Warsaw and Jabłonna (1962), Ed. L. Infeld, Gauthiers-Villars, Paris, 1964; A. Uhlmann, Acta. Phys. Polon. 24, 295, (1963).

69. See refs. 61, 60 and also I. Robinson and A. Trautman, Proc. Roy. Soc. (Lond.) A265, 463 (1962).

70. Due to Demianski (private communication).

71. R.K. Sachs, Proc. Roy. Soc. (Lond.) A264, 309 (1961).

72. J.N. Goldberg and R.P. Kerr, J. Math. Phys. 5, 172 (1964); see also Kozarzewski (ref. 60).

73. For a convenient reference work concerning cosmological questions, see H. Bondi, Cosmology (2nd ed.), Camb. Univ. Press, Cambridge, 1960. See also: E. Schrödinger, Expanding Universes, Camb. Univ. Press, Cambridge, 1956; O. Heckmann and E. Schucking, in Gravitation, Ed. Witten, Wiley,

New York, 1962; H.P. Robertson, Rev. Mod. Phys. 5, 62 (1933); W. Rindler, Mon. Not. Roy. Astr. Soc. 116, 6 (1956); A.G. Walker, J. Lond. Math. Soc. 19, 219 (1944); 19, 227 (1944).

74. R.D. Norman, "Propagation of Relativistic Gravitational Fields" (thesis), London, 1964.

75. S.W. Hawking and G.R.F. Ellis, Phys. Lett. 17, 246 (1965); also C. Behr, Z. Physik, 172, 488 (1963); L.C. Shepley, Proc. Nat. Acad. Sci. 52, 1403 (1964).

76. See, for example, R.W. Lindquist and J.A. Wheeler, Revs. Mod. Phys. 29, 432 (1957).

77. E.M. Lifshitz and I.M. Khalatnokov, Advan. in Phys. 12, 185 (1963).

78. S.W. Hawking, "The Occurrence of Singularities in Cosmology III; Causality and Singularities" (to appear); see also, S.W. Hawking, Proc. Roy. Soc. (Lond.) A294, 511 (1966); A295, 490 (1966).

79. R. Penrose, Phys. Rev. Lett. 14, 57 (1965).

80. See J.A. Wheeler, in Relativity, Groups and Topology, Ed. C. DeWitt and B. DeWitt, Gordon and Breach, New York, 1964.

81. S. Chandrasekhar, M.N. 95, 207 (1935).

82. See, for example, B.K. Harrison, K.S. Thorne, M. Wakano and J.A. Wheeler, Gravitation Theory and Gravitational Collapse, Univ. of Chicago Press, Chicago, 1965; also K.S. Thorne, "Relativistic Stellar Structure and Dynamics" (1966 Les Houches Lectures, to appear) for a general reference.

83. A.S. Eddington, Nature 113, 192 (1924); D. Finkelstein, Phys. Rev. 110, 965 (1958).

84. J.R. Oppenheimer and H. Snyder, Phys. Rev. 56, 455 (1939).

85. M.D. Kruskal, Phys. Rev. 119, 1743 (1960).

86. D.L. Beckedorff and C.W. Misner (unpublished); see ref. 82.

87. D. Finkelstein, lecture at King's College London, 1959. See also ref. 88.

88. R.H. Boyer and R.W. Lindquist, "Maximal Analytic Extension of the Kerr Metric" (to appear in J. Math. Phys. 1967); also B. Carter, Phys. Rev. 141, 1242 (1966)

89. For example, see Quasi-Stellar Sources and Gravitational Collapse, Ed. I. Robinson, A. Schild and E.L. Schücking.

90. J.P. Ostriker, P. Bodenheimer and D. Lynden-Bell, Phys. Rev. Lett. 17, 816 (1966).

91. R.P. Kerr, Phys. Rev. Lett. 11, 237 (1963).

92. P. Jordan, J. Ehlers and R.K. Sachs, Akad. Wiss., Mainz 1, (1961).

93. S. Refsdal, Mon. Not. R. Astr. Soc. 128, 295 (1964); 128, 307 (1964).

94. A.G. Doroshkevich, Ya. B. Zel'dovich and I.D. Novikov (Zhur. Eksp. Teor. Fiz. 49, 170 (1965); Engl. Tr.: Sov. Phys.-JETP, 22, 122 (1966)) have examined slightly aspherical collapse and have verified that trapped surfaces still evolve.

95. The concept of a Cauchy horizon is due to Hawking, see refs. 78.

96. For a comparison between completeness conditions, see W. Kundt, Z. Phys. 172, 488 (1963); M. Fierz and R. Jost, Helv. Phys. Acta 38, 137 (1965).

97. This is a limiting case of the effect of A. Raychauchuri (Phys. Rev. 98, 1123 (1955)) and A. Komar, (Phys. Rev. 104, 544 (1956)).

98. For example, see N. Steenrod, The Topology of Fibre Bundles, Princeton University Press, Princeton, 1951, p. 207.

99. H. Bondi and T. Gold, Mon. Not. R. Astr. Soc. 108, 252 (1948), F. Hoyle and J.V. Narlikar, Proc. Roy. Soc. (Lond.) A273, 1 (1963); A276, 465 (1963); A282, 178 (1964).

100. C.W. Misner, J. Math. Phys. 4, 924 (1963).

101. See, for example, the first-mentioned of refs. 78.

102. See, for example, C. Møller, The Theory of Relativity, Oxford Univ. Press, Oxford, 1952.

103. J.C. Graves and D.R. Brill, Phys. Rev. 120, 1507 (1960).

29
TWISTOR ALGEBRA

J. Math. Phys. **8(2)** (1967), 345–366

This paper initiated twistor theory. It is, however, a little unfortunate that I wrote it using a description of a twistor in terms of a pair of spinors $(\lambda^A,\mu_{A'})$ that I would now consider to be 'the wrong way around'. For a null twistor describing a light ray L in Minkowski space-time, we here have λ^A pointing along L, whose tangent vector n is given by $n^{AA'} = \lambda^A\bar{\lambda}^{A'}$. Instead, we *should* be thinking of λ^A as a *lower*-index quantity (the *secondary*—or *projection*—part of the twistor, where $\mu_{A'}$ ought to be the upper-indexed primary part). This arises because the physical interpretation of n really is a *4-momentum*, and naturally a *co*vector n_a ($n_{AA'} = \lambda_A\bar{\lambda}_{A'}$), which is consistent with behaviour under conformal rescalings. I was not aware of this physical interpretation of a twistor when I wrote this paper, and in any case, not then using abstract indices (see Chapter 36); I would have considered this covariant/contravariant distinction not to matter much. The first publication bringing out the momentum and angular-momentum interpretation of a twistor is Chapter 46 (explicitly, only in the footnote on p. [28]). The now-standard $(\omega^A, \pi_{A'})$ notation (where $\omega^A = -\bar{\mu}^A$ and $\pi_{A'} = \bar{\lambda}_{A'}$) does not appear explicitly until Chapter 52.

Twistor Algebra

R. Penrose
Department of Mathematics, Birkbeck College, London, England
(Received 16 May 1966)

A new type of algebra for Minkowski space-time is described, in terms of which it is possible to express any conformally covariant or Poincaré covariant operation. The elements of the algebra (twistors) are combined according to tensor-type rules, but they differ from tensors or spinors in that they describe locational properties in addition to directional ones. The representation of a null line by a pair of two-component spinors, one of which defines the direction of the line and the other, its moment about the origin, gives the simplest type of twistor, with four complex components. The rules for generating other types of twistor are then determined by the geometry. One-index twistors define a four-dimensional, four-valued ("spinor") representation of the (restricted) conformal group. For the Poincaré group a skew-symmetric metric twistor is introduced. Twistor space defines a complex projective three-space C, which gives an alternative picture equivalent to the Minkowski space–time M (which must be completed by a null cone at infinity). Points in C represent null lines or "complexified" null lines in M; lines in C represent real or complex points in M (so M, when complexified, is the Klein representation of C. Conformal transformations of M, including space and time reversals (and complex conjugation) are discussed in detail in twistor terms. A theorem of Kerr is described which shows that the complex analytic surfaces in C define the shear-free null congruences in the real space M. Twistors are used to derive new theorems about the real geometry of M. The general twistor description of physical fields is left to a later paper.

I. INTRODUCTION

In the study of Lorentz covariant local field theories, field quantities (with spin) are normally described by vectors, tensors, or, more generally, by spinors, all finite-dimensional representations of the Lorentz group[1] being expressible in terms of spinors (vectors and tensors being regarded as effectively special cases). A local (restricted) Lorentz transformation then takes the form

$$\psi_{E\cdots J'\cdots}^{A\cdots G'\cdots} \to \psi_{Q\cdots V'\cdots}^{M\cdots S'\cdots} t_M^A \cdots \bar{t}_S^{G'} t^{(-1)Q}_{\ \ E} \cdots \bar{t}^{(-1)V'}_{\ \ J'}, \tag{1.1}$$

(t_M^A) being a complex unimodular (2×2) matrix, with inverse $(t^{(-1)A}_{\ \ M})$. Except where general relativity is involved, normally it is further required that physical quantities be suitably covariant under the full Poincaré group.[1] However, this covariance is expressed (locally) in a totally different way from the local Lorentz covariance. Whereas the dependence on space–time *direction* is expressed *algebraically* [cf., (1.1)], the dependence on *position* is described by *differential* equations (so that the analog of (1.1) involves integrals). The invariance of all expressions under the translation

$$x^i \to x^i + a^i \tag{1.2}$$

is generally ensured by the fact that the (Minkowskian) coordinates x^i must, themselves, never enter into the field equations explicitly, only the operators $\partial/\partial x^i$ being permitted to occur.

It is curious that the invariance under two types of transformation (1.1) and (1.2), each of which simply refers to symmetries of the Minkowski space–time, should find mathematical expression in so different a way. This is, moreover, *not* just a property of a difference in group structure between the translation and rotation elements of the Poincaré group, but rather a consequence of *what is meant by a local field theory*. In fact, mathematical formalisms do exist[2] in which *all* the Poincaré transformations are represented according to an algebraic "tensor type" law similar to (1.1). It is the purpose of the present paper to exhibit and discuss in some detail one such formalism, the basic elements of which are re-

[1] Throughout this paper, "Lorentz group" always refers to to the six-parameter *homogeneous* group and "Poincaré group" to the ten-parameter *inhomogeneous* group. "Conformal group" refers to the fifteen-parameter group of transformations which preserve the local conformal structure of Minkowski space–time. If in any particular context it is important to exclude the space reversing and time reversing transformations, then this is made explicit, for example, by the use of the term "restricted".

[2] The use of homogeneous coordinates in space–time would afford a simple example of such a formalism (but apparently one of limited physical interest). More significant, of course, is the representation of physical quantities in terms of Hilbert space, which has the advantage that the (infinite demensional) analogs of $(t_M^{\ A})$ are unitary [cf., particularly E. P. Wigner, Ann. Math. 40, 149 (1939) V. Bargmann and E. P. Wigner, Proc. Natl. Acad. Sci. U. S. 34, 211 (1948)]. However, it will be essential here to preserve the *finite* dimensionality of the operations at *this* stage, so that geometrical questions can be kept in the forefront.

ferred to here as *twistors*. It will turn out that the twistor algebra will have the same type of universality, in relation to the *conformal* group,[1] that the well-known and highly effective two-component spinor algebra of van der Waerden[2] has, in relation to the Lorentz group. Twistors are, in fact, the "spinors" which are relevant to the six-dimensional space whose (pseudo-) rotation group is isomorphic with the conformal group of ordinary Minkowski space–time.[4] The simplest (non-scalar) twistors constitute a *four-dimensional, four-valued representation of the restricted conformal group*[5] (eight-dimensional if reflections are included). The general twistor is then a many-index quantity constructible from the above basic twistors by means of the usual "tensor type" rules. Any finite-dimensional representation of the conformal group is thus expressible as a direct sum of twistor representations.[4]

The emphasis here will be on the *geometrical* aspects of twistors. It will, in fact, be possible to give a fairly complete geometrical picture of twistors and of their basic operations. Ordinary space–time concepts can then be translated into twistor terms. However, the geometrical expressions of the most immediate twistor concepts have a somewhat nonlocal character. Thus, the primary geometrical object will not be a point in Minkowski space–time, but rather a null straight line or, more generally, a twisting congruence of null lines. Points do, in fact, emerge, but only at a secondary stage. (It also turns out that a natural description of physical fields in twistor terms is given by quantities having a nonlocal space–time interpretation.) However, any vector, tensor, or spinor operation *can* be translated into twistor terms, if desired, and *vice versa*.

All the basic operations of the formalism are conformally invariant. This is of relevance in the study of zero rest-mass fields, since, for each spin, the free-field equations are all effectively conformally invariant.[6,7] This latter fact has particular importance in the asymptotic analysis of such fields,[7] and it would appear that a "manifestly conformally invariant" formalism should be valuable to this type of analysis. Furthermore, although the presence of *mass* breaks conformal invariance, the conformal group appears to have relevance as an approximate symmetry in very high energy physics.[8] When the energies of the particles are high enough, their rest masses can be neglected.

However, the applicability of the twistor formalism is not restricted to situations in which the conformal group is relevant. Since the Poincaré group is a subgroup of the conformal group, it follows that Poincaré covariant operations can be expressed in twistor terms. It merely becomes necessary to introduce a fixed (skew-symmetric) "metric twistor" $I^{\alpha\beta}$ which singles out a particular Minkowski structure (consistent with the given conformal structure). By including operations involving $I^{\alpha\beta}$ with the basic twistor operations, a formalism invariant under the Poincaré group is obtained, in terms of which, any Poincaré covariant operation is, in principle, expressible. Furthermore, by making slightly different alternative choices for $I^{\alpha\beta}$, the corresponding formalisms for a class of space–times including the de Sitter and Einstein cosmologies are obtained.

Perhaps the most obvious drawback of the twistor formalism, however, from the point of view of a possible *fundamental* applicability in physics, is that it is so tied to the idea of a *conformally flat* background space–time, that it is difficult to conceive of how to incorporate the formalism (as it stands) completely into general relativity. It is not impossible that some modification of twistor algebra might be applicable to general curved background space–times. Indeed, some of the basic ideas of the formalism, namely those concerning the null line congruences (cf., the results of Robinson and Kerr referred to in Sec. VIII) have, as their origins, some researches into general relativity. However, any such modification would have to be of a different order from the comparatively straightforward adaption of spinor algebra into general relativity.[9] The point of view one must apparently take, is that whereas the vector–tensor–spinor algebra refers to the im-

[1] B. L. van der Waerden, Nachr. Ges. Wiss. Gottingen **100**, 1 (1929).

[4] F. Klein, *Gesammelte Mathem. Abhandlungen* (J. Springer, Berlin, 1921); cf. also H. Weyl, *The Classical Groups* (Princeton University Press, Princeton, New Jersey, 1939); R. Brauer and H. Weyl, Am. J. Math. **57**, 425 (1935).

[5] Quantities which are essentially twistors have been described by W. A. Hepner, Nuovo Cimento **26**, 351 (1962). See also Y. Murai, Nucl. Phys. **6**, 489 (1958); Progr. Theoret. Phys. (Kyoto) **9**, 147 (1953); **11**, 441 (1954). Similar quantities have also recently gained prominence in, for example, the work of A. Salam, R. Delbourgo, and J. Strathdee, Proc. Roy. Soc. (London) **A284**, 147 (1965).

[6] E. Cunningham, Proc. London Math. Soc. **8**, 77 (1910); H. Bateman, *ibid.* **8**, 223 (1910); P. A. M. Dirac, Ann. Math. **37**, 429 (1936); J. A. McLennan, Jr., Nuovo Cimento **10**, 1360 (1956); H. A. Buchdahl, *ibid.* **11**, 496 (1959).

[7] R. Penrose, Proc. Roy. Soc. (London) **A284**, 159 (1965).

[8] H. A. Kastrupp, Phys. Letters (Amsterdam) **3**, 78 (1962); Phys. Rev. **142**, 1060 (1966). A recent additional suggestion is that the mass splittings of strong interaction physics may be derivable from conformal group symmetry: D. Bohm, M. Flato, D. Sternheimer, and J. P. Vigier, Nuovo Cimento **38**, 1941 (1965). For a discussion of the relevance of the conformal group in physics, see T. Fulton, F. Rohrlich, and L. Witten, Rev. Mod. Phys. **34**, 442 (1962).

[9] L. Infeld and B. L. van der Waerden, S. B. Preuss. Akad. **9**, 380 (1933).

mediate neighborhood of a "point" in space–time, the twistor algebra refers to a more extended "nonlocal" region. It seems to be mathematically more difficult to piece together such regions into a curved space–time, than to build a space–time out of points.

Finally, one of the most important initial motivations for considering a formalism of this kind should be briefly mentioned, although implications of this idea are not discussed in this paper. It concerns possible applications to quantum field theory. One of the most significant quantum mechanical operations is the splitting of field amplitudes into their positive and negative frequency parts. Although this is generally expressed in terms of Fourier transforms, there is the alternative description in terms of singularity-free analytic extensions into the upper or lower complex half-planes.[10] In the present formalism, the idea is to use this second approach and to combine it with global properties of complexified Minkowski space–time. Similarly to the way that the real axis divides the complex plane into two disconnected halves, the space of "null" twistors (describing real Minkowskian null lines) divides twistor space into two disconnected halves, namely the "right-handed" and the "left-handed" twistors (which may be thought of as describing "complexified" null lines).

Free fields of zero rest mass and arbitrary spin can be described particularly conveniently in twistor terms. Moreover, it turns out that they can be generated in a remarkably simple way as contour integrals of arbitrary analytic functions defined in twistor space. (A *null* field is generated when the contour surrounds a simple pole of the function. This leads to a certain generalization of a theorem of Robinson.[11]) If the relevant singularities of this function are associated with one half, rather than the other, of the twistor space, this ensures that the field is of (say) positive frequency. Other properties of zero rest-mass fields also find a natural expression in twistor terms, notably the interrelation with certain types of potential fields[7] and the structure of their total energy-momentum–angular-momentum (which give ten of the fifteen components of a trace-free "Hermitian" twistor $E^{\alpha}_{\dot{\beta}}$). The discussion of these matters is left to a later paper.

Care has been taken here to illustrate that twistor algebra is not merely an abstract formalism, but that the algebraic operations have well-defined meanings in terms of space–time geometry. For whether or not twistors have a significant role to play in future physical theory, the results so far obtained suggest strongly that the formalism ought at least to be thoroughly explored from both geometric and analytic points of view.

[10] See any standard work on quantum field theory. The essential matters referred to here can be found in R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (W. A. Benjamin, Inc., New York, 1964).

[11] I. Robinson, J. Math. Phys. **2**, 290 (1961).

II. TWISTORS AND NULL LINES

As a starting point for the geometrical description of a twistor, consider a null straight line L in Minkowski space–time M. Choose a set of Minkowski coordinates[12] (x^i) for M with origin O. Let l^i be the position vector of some point P on L and let n^i be a future-pointing tangent vector to L (see Fig. 1). If we wish to assign a set of coordinates to the line L, we may do this (Plücker–Grassmann coordinates) by selecting n^i and the moment

$$m^{ij} = l^i n^j - n^i l^j \tag{2.1}$$

of the vector n^i (acting at P) about O. Then, the ratios of the ten quantitites (n^i, m^{ij}) will uniquely define L. (Note that this is independent of the choice of P on L.) This is, however, a highly redundant representation. In addition to the requirement here that n^i be null

$$n^i n_i = 0, \tag{2.2}$$

there are the consistency relations for (2.1)

$$\epsilon_{ijkl} m^{ij} n^k = 0 \tag{2.3}$$

[and also $\epsilon_{ijkl} m^{ij} m^{kl} = 0$, which is implied by (2.3)]. Equation (2.3) in fact represents just three independent conditions which, together with (2.2), reduce the set of nine ratios in (n^i, m^{ij}) to just *five* independent real numbers. This is consistent with

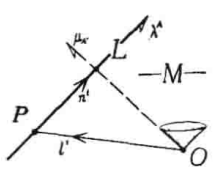


FIG. 1. The representation of a null line.

[12] World vectors and tensors are labeled by lower case Latin indices running over 0,1,2,3, the Minkowski metric being given by $(g^{ij}) = \mathrm{diag}(1,-1,-1,-1)$. Capital Latin index letters, primed or unprimed, denote spinor indices and run over 0,1 or 0′,1′. Greek indices are twistor indices and run over 0,1,2,3. The summation convention applies to each of these four types of index separately. Thus, in particular, no summation takes place between primed and unprimed spinor indices even when the same letter is used, e.g., J and J' are regarded as distinct letters in $x_J{}^{J'}$. This allows us to write the tensor–spinor correspondence in a definite way by simply using the two capital versions of a tensor index as its spinor translation: $x^i \leftrightarrow x^{JJ'}$, etc. (Rindler's convention).

the fact that the null lies in M form an ∞^5 system (for, choosing P as the intersection of L with a fixed spacelike hyperplane, we have ∞^3 choices for P and ∞^2 choices for the null direction at P).

So far, the fact that L is null has not played an essential part. But null directions are very conveniently represented in terms of *two-component spinors*. It will emerge that, by using spinors, we can greatly simplify the description of the null line L; the set of four conditions (2.2), (2.3) being effectively replaced by a *single* condition. Since the following discussion depends essentially on the use of spinors, a very brief review of the ideas required will now be given.[13] The translation from world tensors to spinors is achieved using a quantity[12] $\sigma_i^{JJ'}$ [a Hermitian (2 × 2) matrix for each j] and its inverse $\sigma^i_{JJ'}$, subject to $\sigma_i^{JJ'}\sigma_k^{KK'}\epsilon_{JK}\epsilon_{J'K'} = g_{ik}$, $\sigma_i^{JJ'}\sigma^i_{JJ'} = \delta^i_j$. The ϵ_{JK}, $\epsilon_{J'K'}$ are skew-symmetric Levi-Civita symbols and are used for raising and lowering spinor indices: $\xi^A\epsilon_{AB} = \xi_B$, $\epsilon^{AB}\xi_B = \xi^A$, i.e., $\xi^0 = \xi_1$, $\xi^1 = -\xi_0$, and similarly for primed indices. Any tensor (e.g., $X^{ij}{}_k$) has a spinor translation, which is written using the same base symbol, but with each tensor index replaced by the corresponding pair of spinor indices, e.g.,

$$X^{ij}{}_k \leftrightarrow X^{II'JJ'}{}_{KK'} = X^{ij}{}_k\sigma_i^{II'}\sigma_j^{JJ'}\sigma^k{}_{KK'}.$$

Under complex conjugation, the roles of primed and unprimed indices are interchanged, so that reality of tensors is expressed as Hermiticity of spinors. Since M is flat, we can here choose $\sigma_i^{JJ'}$ constant and equal to $2^{-\frac{1}{2}}$ times the unit matrix and Pauli matrices. Then, say,

$$(x^0, x^1, x^2, x^3) \leftrightarrow \begin{pmatrix} x^{00'} & x^{01'} \\ x^{10'} & x^{11'} \end{pmatrix} = 2^{-\frac{1}{2}}\begin{pmatrix} x^0 + x^1 & x^2 + ix^3 \\ x^2 - ix^3 & x^0 - x^1 \end{pmatrix}. \tag{2.4}$$

The spinor translation of a complex null vector α^i (i.e., $\alpha^i\alpha_i = 0$) has the form

$$\alpha^i \leftrightarrow \alpha^{JJ'} = \beta^J\gamma^{J'}. \tag{2.5}$$

If α^i is real and future pointing, then we can take $\gamma^{J'}$ to be the complex conjugate of β^J, i.e.,

$$\alpha^i \leftrightarrow \beta^J\bar{\beta}^{J'}. \tag{2.6}$$

Equation (2.6) implies a geometrical realization of a spinor β^J, up to a phase multiplier, namely as a future-pointing null vector. Moreover, β^J can be completely realized geometrically[14] up to *sign*, in terms of the *bivector*

$$\pi^{ik} \leftrightarrow \beta^J\beta^K\epsilon^{J'K'} + \epsilon^{JK}\bar{\beta}^{J'}\bar{\beta}^{K'}. \tag{2.7}$$

This is real (and null), and it determines a *half-plane element* (the "flag plane") tangent to the light cone along the vector α^i. However, the sign of β^J cannot be realized *locally*[15] geometrically, since a continuous rotation through 2π changes β^J into $-\beta^J$. If we are interested in β^J only up to *proportionality*, then the complete geometrical realization is simply as a null *direction*.

Let us now represent, in spinor terms, the quantities n^i, m^{ij}, which define the null line L. We have

$$n^i \leftrightarrow n^{JJ'} = \lambda^J\bar{\lambda}^{J'} \tag{2.8}$$

and, from (2.1),

$$m^{ik} \leftrightarrow l^{JJ'}\lambda^K\bar{\lambda}^{K'} - \lambda^J\bar{\lambda}^{J'}l^{KK'} = i\epsilon^{JK}\mu^{(J'}\bar{\lambda}^{K')} - i\mu^{(J}\lambda^{K)}\epsilon^{J'K'}, \tag{2.9}$$

where

$$\mu_{A'} = -i\lambda^A l_{AA'}. \tag{2.10}$$

(Round brackets denote symmetrization. The factor $-i$ is for later convenience.) Thus λ^A and $\mu_{A'}$ together determine l^i and m^{ik}. We may think of λ^A as defining the *direction* of L and $\mu_{A'}$ as effectively giving us the *moment* of λ^A ("acting" at P) about O. It is clear from (2.10) that, if λ^A is multiplied by any complex factor, then L is unchanged if $\mu_{A'}$ is multiplied by the same factor. Furthermore, (2.10) implies that $\mu_{A'}$ is independent of the choice of P on L (i.e., if $l_{AA'} \to l_{AA'} + a\lambda_A\bar{\lambda}_{A'}$, then $\mu_{A'}$ is unchanged since $\lambda^A\lambda_A = 0$).

Note a particular choice of P which is of interest, namely the intersection of L with the null cone of O. (This intersection exists uniquely provided L does not lie in any null hyperplane through O.) Then l^i or $-l^i$ has the form (2.6) and it follows from (2.10) that

$$l_{AA'} = i(\lambda^B\bar{\mu}_B)^{-1}\bar{\mu}_A\mu_{A'}. \tag{2.11}$$

[13] For more complete accounts see Refs. 3 and 9 and, for example, W. L. Bade and H. Jehle, Rev. Mod. Phys. **25**, 714 (1953); E, M. Corson, *An Introduction to Tensors, Spinors and Relativistic Wave Equations* (Blackie & Son Ltd., London, 1953); F. A. E. Pirani, in *Brandeis Summer Institute in Theoretical Physics*, 1964, *Lectures on General Relativity* (Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1965), Vol. 1.

[14] E. T. Whittaker, Proc. Roy. Soc. (London) **A158**, 38 (1937); W. T. Payne, Am. J. Phys. **20**, 253 (1952); R. Penrose, "Null Hypersurface Initial Data", in P. G. Bergmann's Aeronautical Research Lab. Tech. Documentary Rept. 63–56 (Office of Aerospace Research, U. S. Air Force, 1963).

[15] To give a rigorous definition of a spinor which takes into account its sign, it is usual to appeal to the theory of fibre bundles. This is not essential, however, and an elementary (nonlocal) geometrical description will be given in an appendix to a forthcoming book by R. Penrose and W. Rindler on the applications of spinors in relativity.

Thus, the null direction defined by $\mu_{A'}$ [cf., (2.6)] is that of the null line through O which meets L (Fig. 1). The exceptional case $\lambda^B \bar{\mu}_B = 0$ corresponds to L lying in a null hyperplane through O. [This follows from (2.10), since n^i would be necessarily orthogonal to any choice of l_i.] In this case, λ^A and $\bar{\mu}^A$ are proportional, so the null direction of $\mu_{A'}$ is that of L. More exceptionally still, L might pass through O. This is the case $\mu_{A'} = 0$. On the other hand, for the present, we do not allow $\lambda^A = 0$ (although this will be reconsidered shortly).

The null line L can now be assigned, for coordinates, the three complex ratios of the four complex quantities

$$L^0 = \lambda^0, \qquad L^1 = \lambda^1, \qquad L^2 = \mu_{0'}, \qquad L^3 = \mu_{1'}, \tag{2.12}$$

which we write

$$(L^\alpha) = (\lambda^A, \mu_{A'}). \tag{2.13}$$

These three complex ratios are equivalent to *six* real parameters, so we must expect to find one real relation connecting λ^A and $\mu_{A'}$. This is obtained from (2.10), since the *reality* of the vector l^i implies $l_{AA'}$ is Hermitian, whence

$$\mathrm{Re}\,(\lambda^A \bar{\mu}_A) = 0. \tag{2.14}$$

Provided $\lambda^A \neq 0$, condition (2.14) is also *sufficient* to ensure the existence of a null line L associated with λ^A and $\mu_{A'}$. For, if $\lambda^A \bar{\mu}_A$ is pure imaginary and nonvanishing, then (2.11) gives us a point P, through which we choose L with direction given by λ^A. On the other hand, if $\lambda^A \bar{\mu}_A = 0$, we can easily solve (2.10) to obtain P, with OP spacelike ($l_{AA'}$ of the form $\lambda_A \bar{\nu}_{A'} + \nu_A \bar{\lambda}_{A'}$).

A quantity with components (L^α) given as in (2.13) will be called a *null twistor* (of valence $\left[{}^1_0\right]$) if (2.14) holds.[16] We shall also be interested in such quantities when (2.14) does *not* hold (and also when λ^A is allowed to vanish). We may regard such a quantity as describing, in some sense, a "complexified" null line [when (2.14) fails], since if we allow l^i to be a *complex* vector, then (2.14) would, in general, be violated. [This is not a straightforward complexification of the null lines in M in the ordinary sense, however, since the real dimensionality of the system of lines is only increased from five to six by dropping (2.14); but cf., Sec. VI.] If $\mathrm{Re}\,(\lambda^A \bar{\mu}_A) > 0$, we refer to L^α as a *right-handed twistor* (valence $\left[{}^1_0\right]$) and if $\mathrm{Re}\,(\lambda^A \bar{\mu}_A) < 0$, a *left-handed twistor*. The reason for this terminology will become evident later. When $\lambda^A = 0$, we shall still refer to L^α as a *null* twistor even though no finite null line L is represented. The role of such twistors will emerge shortly.

III. INCIDENCE OF NULL LINES IN TWISTOR TERMS

The algebraic rules for the manipulation of twistors will have, as their basis, the idea of *incidence* between null lines. Let us consider, in addition to L, a second null line X, described by the null twistor X^α:

$$(X^\alpha) = (\xi^A, \eta_{A'}), \tag{3.1}$$

where ξ^A and $\eta_{A'}$ are the analogs of λ^A, $\mu_{A'}$ above. That is, ξ^A defines the direction of X and we have

$$\eta_{A'} = -i\xi^A x_{AA'}, \tag{3.2}$$

where x^i is the position vector of a point on X (see Fig. 2). Suppose now that X and L do intersect. Then we may choose $l^i = x^i$ for the coordinate vector of this intersection point, whence by (2.10) and (3.2),

$$\xi^A \bar{\mu}_A = i\xi^A l_{AA'} \bar{\lambda}^{A'} = i\xi^A x_{AA'} \bar{\lambda}^{A'} = -\eta_{A'} \bar{\lambda}^{A'}. \tag{3.3}$$

Let us define the *complex conjugate* of a twistor L^α to be $\bar{L}_\alpha$ (valence $\left[{}^0_1\right]$), where

$$(\bar{L}_\alpha) = (\bar{\mu}_A, \bar{\lambda}^{A'}) \tag{3.4}$$

when L^α is given by (2.13) [irrespective of the condition (2.14), or whether or not λ^A vanishes]. In component form[17]:

$$\bar{L}_0 = \overline{L^2}, \qquad \bar{L}_1 = \overline{L^3}, \qquad \bar{L}_2 = \overline{L^0}, \qquad \bar{L}_3 = \overline{L^1}. \tag{3.5}$$

Then, (3.3) tells us that a necessary condition for

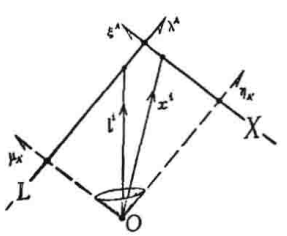


FIG. 2. Incidence of two null lines.

[16] Any temptation to identify the twistor (2.13) with a Dirac spinor should be rejected here, since their transformation properties are quite different [cf., for example, (7.13), (7.17)].

[17] The operation of twistor (or spinor) complex conjugation is denoted by a bar extending only over the base symbol involved and not over the indices. If the bar extends also over the indices, this denotes simply the complex conjugate of the complex number that the symbol represents.

the null lines L and X to meet, in terms of their corresponding twistors, is

$$X^a\bar{L}_a = 0, \tag{3.6}$$

since

$$X^a\bar{L}_a = \xi^A\bar{\mu}_A + \eta_{A'}\bar{\lambda}^{A'}. \tag{3.7}$$

Note, further, that the condition (2.14) for L^a to represent a real null line is simply

$$L^a\bar{L}_a = 0. \tag{3.8}$$

We can think of (3.8) as a special case of (3.6) since L intersects itself! We also have, here,

$$X^a\bar{X}_a = 0. \tag{3.9}$$

Condition (3.6) is, conversely, also *sufficient* for the null lines L and X to intersect, where we assume, for the moment, that λ^A and ξ^A are not proportional, so that L and X are not parallel. For, we then have $\xi^A\lambda_A \neq 0$ and we can construct the (complex) vector

$$p_j \leftrightarrow p_{JJ'} = (i/\xi^A\lambda_A)(\lambda_J\eta_{J'} - \xi_J\mu_{J'}). \tag{3.10}$$

We observe that

$$\mu_{A'} = -i\lambda^A p_{AA'}, \qquad \eta_{A'} = -i\xi^A p_{AA'}.$$

Thus, when p_j is *real*, we can satisfy (2.10) and (3.2) by putting $l_j = p_j = x_j$, whence L and X must intersect. (In fact, it is valid to regard the "complex point" with position vector p^j as the "intersection" of L and X even in the cases when p_j is not real, as we shall see later.) Now p_j is real if and only if $p_{JJ'}$ is Hermitian. Since λ^A and ξ^A are not proportional, we can test the Hermiticity of $p_{JJ'}$ by taking components with respect to λ^A, ξ^A:

$$\lambda^A\bar{\lambda}^{A'}(p_{AA'} - \bar{p}_{A'A}) = i\mu_{A'}\bar{\lambda}^{A'} + i\lambda^A\bar{\mu}_A = iL^a\bar{L}_a,$$

$$\xi^A\xi^{A'}(p_{AA'} - \bar{p}_{A'A}) = i\eta_{A'}\xi^{A'} + i\xi^A\bar{\eta}_A = iX^a\bar{X}_a,$$

$$\xi^A\bar{\lambda}^{A'}(p_{AA'} - \bar{p}_{A'A}) = i\eta_{A'}\bar{\lambda}^{A'} + i\xi^A\bar{\mu}_A = iX^a\bar{L}_a.$$

Thus, the reality of p_j is a consequence of (3.8), (3.9), and (3.6), so that in the case of twistors representing nonparallel null lines, we can state the condition that the lines meet simply as the orthogonality condition (3.6) between the twistors.

We may further ask what is the geometrical meaning of this condition $X^a\bar{L}_a = 0$ when X and L are real *parallel* null lines. Perhaps the simplest way to see the result is to use a limiting argument. Keep L_a fixed but vary X^a, so that $X^a\bar{L}_a = 0$ throughout the motion. Suppose L and X are initially not parallel, but become parallel in the limit. The intersection point recedes to infinity along L. The possible positions of X, for each *finite* position of the intersection point, are the generators of the null cone of this point. As the point recedes to infinity, the null cone becomes, in the limit, the null hyperplane through X. Thus, when L and X are parallel, the condition $X^a\bar{L}_a = 0$ becomes the condition that *L and X lie in the same null hyperplane.*

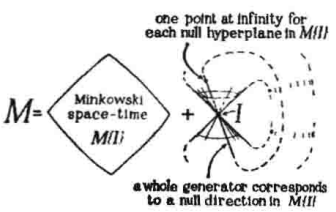

one point at infinity for each null hyperplane in M{I}
M=
Minkowski space-time M{I}
+
I
a whole generator corresponds to a null direction in M{I}

FIG. 3. The construction of the compact manifold $\boldsymbol{M}$.

In fact, it is convenient to regard all the null lines of a null hyperplane as intersecting in a single *point at infinity*. Thus, we adjoin to our space M a set of ∞^3 such points at infinity, one for each of the null hyperplanes in M. The geometrical role played by the twistors L_a with $\lambda^A = 0$ (but $\mu_{A'} \neq 0$) now emerges. For, if $\lambda^A = 0$, the condition $X^a\bar{L}_a = 0$ becomes $\xi^A\bar{\mu}_A = 0$ [cf., (3.7)]. That is to say, *the direction of X coincides with the null direction represented by $\mu_{A'}$*, so that for *fixed* L^a the corresponding lines X are all parallel. Each null hyperplane of parallel lines X gives rise to *one* point at infinity; the aggregate of all these points at infinity, corresponding to all these parallel hyperplanes, gives an ∞^1 system of points at infinity, namely the points of the *null line at infinity* L. In fact, to complete the picture, we must add one further point at infinity, not lying on any finite line, which we call I. The point I is common to all the lines L at infinity. The fact that any two null lines at infinity must, for consistency here, be considered to intersect, follows from (3.7), since if $\lambda^A = 0$ and also $\xi^A = 0$, then $X^a\bar{L}_a = 0$ automatically follows. The point I then plays the part of the *vertex* of a *null cone at infinity* the generators of which are the lines at infinity considered above (see Fig. 3).

The structure of the completed ("compactified") Minkowski space arrived at in this way, by adding a (closed) null cone at infinity, is one which has been considered by a number of authors.[18,7] We henceforth use the symbol M, here, to refer to the *entire* completed space and not just to the set of finite points. The set of finite points is instead denoted by $M\{I\}$ to indicate that the null cone of the

[18] N. H. Kuiper, Ann. Math. 50, 916 (1949); H. Rudberg, dissertation, University of Uppsala, Uppsala, Sweden (1958); R. Penrose, in *Proceedings of the 1962 Conference on Relativistic Theories of Gravitation, Warsaw* (Polish Academy of Science, Warsaw, 1965) A. Uhlmann, Acta Phys. Polon. 24, 293 (1963). Rudberg also mentions the four-valuedness of spinors.

particular point I has been removed from M. The geometry of M is briefly as follows: M is a compact nonsingular manifold without boundary which has a well-defined conformal structure everywhere (i.e., a metric defined up to proportionality). It has a transitive ∞^{15} group of motions preserving this conformal structure, every point being on an equal footing with every other point. M contains ∞^5 null "straight" lines (geodesics) each of which is topologically a circle. Through each point P of M pass ∞^2 of these lines generating the (closed up) null cone of P. The removal of any one of these null cones leaves a space with Euclidean topology which can be assigned a Minkowskian metric (unique up to dilatations) consistent with the given conformal structure. The entire space M has the topology $S^3 \times S^1$ and can be realized as a real projective quadric fourfold with signature $(+\,+\,-\,-\,-\,-)$.

The null lines in M are in one-to-one correspondence with the proportionality classes of twistors L^α (i.e., κL^α represents the same line as L^α, $\kappa \neq 0$) which satisfy $L^\alpha \bar{L}_\alpha = 0$ and $L^\alpha \neq 0$. Incidence between null lines in M is expressed as orthogonality ($X^\alpha \bar{L}_\alpha = 0$) between the corresponding twistors.

IV. ROBINSON CONGRUENCES

Thus far, the geometrical description of a twistor given here has been restricted to the case of null twistors (of valence $[^1_0]$). As mentioned earlier, if on the other hand $L^\alpha \bar{L}_\alpha \neq 0$, then we may think of L^α as representing, in some sense, a kind of "complexified" null line L. This is, in fact, the way we tend to view "geometrically" a general twistor of valence $[^1_0]$. However, it is significant that a much more precise realization can be given in terms of congruences[19] of null lines. In order to fix our ideas, we regard such a congruence as representing, rather, a twistor of valence $[^0_1]$ (e.g., $\bar{L}_\alpha$), although this makes little difference in practice.

Now, any (real) null line L in M can be completely characterized by *the system of all null lines which meet it*. In twistor terms, that is, if we know the set of all X^α satisfying $X^\alpha \bar{L}_\alpha = 0$, $X^\alpha \bar{X}_\alpha = 0$ (where $L^\alpha \bar{L}_\alpha = 0$, $L^\alpha \neq 0$), then we know L^α up to proportionality. Thus, while we think of L^α as describing the null line L in M, we can think of $\bar{L}_\alpha$ as describing the *congruence* L of null lines (i.e., a three-dimensional system[19] of lines in M) which meet L. In exactly the same way we can represent a general twistor R_α of valence $[^0_1]$, in terms of the congruence R of lines X whose twistors X^α (with $X^\alpha \bar{X}_\alpha = 0$) satisfy $X^\alpha R_\alpha = 0$. If $R_\alpha \bar{R}^\alpha \neq 0$, such a congruence is referred to here as a *Robinson congruence.*[20] [If $R^\alpha \bar{R}^\alpha = 0$, then we denote by $\bar{R}$ the line which is met by all the lines of the congruence R. On the other hand, if L denotes the line, then $\bar{L}$ denotes the corresponding congruence. In general, we adopt the convention that the operation of applying a bar to any aymbol is its own inverse. From this it follows that the definition of $\bar{R}^\alpha$, given R_α, exactly mirrors the definition (3.5) of $\bar{L}_\alpha$, given L^α, since we could put $\underline{L^\alpha = \bar{R}^\alpha}$. Thus,[17] $(\bar{R}^\alpha) = (\bar{R}^0, \bar{R}^1, \bar{R}^2, \bar{R}^3) = (\overline{R_2}, \overline{R_3}, \overline{R_0}, \overline{R_1})$.] If $R_\alpha \bar{R}^\alpha > 0$, we refer to the Robinson congruence R as right-handed. If $R_\alpha \bar{R}^\alpha < 0$, it is left-handed.

Since Robinson congruences occupy such a basic position in the geometry of twistors, it seems worthwhile to examine a particular such congruence in some detail here. (In fact, any particular case is completely representative of the general case, since it turns out that any two Robinson congruences can be transformed one into the other by a Poincaré transformation, perhaps involving a reflection.)

Choose a real number ϵ and put

$$(R_\alpha) = (\epsilon/\sqrt{2},\ 0,\ 1,\ 0) \tag{4.1}$$

so that

$$R_\alpha \bar{R}^\alpha = \epsilon\sqrt{2}. \tag{4.2}$$

Let X^α be as in (3.1): $(X^\alpha) = (\xi^0,\ \xi^1,\ \eta_{0'},\ \eta_{1'})$, so that the condition for the line X to belong to the conrgruence R is

$$\xi^0 \epsilon/\sqrt{2} + \eta_{0'} = 0. \tag{4.3}$$

Write the position vector x^j of a point on X as

$$x^0 = t, \qquad x^1 = z, \qquad x^2 = x, \qquad x^3 = y.$$

Equation (3.2), with (2.4), gives

$$(\eta_{0'},\ \eta_{1'}) = -(i/\sqrt{2})(\xi^0,\ \xi^1)\begin{pmatrix} t - z & -x + iy \\ -x - iy & t + z \end{pmatrix}.$$

Using (4.3), we then get $-\epsilon\xi^0 = i\xi^0(-t + z) + i\xi^1(x + iy)$, i.e.,

$$\xi^0 : \xi^1 = x + iy : t - z + i\epsilon. \tag{4.4}$$

Now $\xi^{J}\bar{\xi}^{J'}$ corresponds to a tangent vector to X, so that

$$\begin{pmatrix} \xi^0\bar{\xi}^{0'} & \xi^0\bar{\xi}^{1'} \\ \xi^1\bar{\xi}^{0'} & \xi^1\bar{\xi}^{1'} \end{pmatrix} \propto \begin{pmatrix} dt + dz & dx + i\,dy \\ dx - i\,dy & dt - dz \end{pmatrix}.$$

[19] The term "congruence" is used to denote a system of curves (or surfaces, etc.) for which there is just one member of the system (or at most a discrete number) through a general point in the space.

[20] First investigated by I. Robinson (private communication).

Combining this with (4.4) we get the differential equations for the line X

$$dt + dz : dx - i\,dy = dx + i\,dy : dt - dz = x + iy : t - z + i\epsilon. \quad (4.5)$$

The solution of this is

$$\begin{aligned} t - z + i\epsilon &= (x + iy)\alpha, \\ (x - iy) - (t + z)\alpha &= \beta, \end{aligned} \quad (4.6)$$

where α and β are complex constants defining the particular line X of R, for consistency satisfying

$$\mathrm{Im}\,(\bar{\beta}\alpha) = \epsilon. \quad (4.7)$$

In order to get a more visualizable picture of the situation, let us consider the intersection with a spacelike hyperplane $t = \tau$ (τ const). Each null line X meets this hyperplane in one point (τ, x, y, z). The direction of X there can be represented by considering the projection into the hyperplane of the tangent vector to X at this point. These projected tangent vectors are then tangents to a series of curves in the hyperplane which should give some kind of a picture of the structure of the Robinson congruence. To find the differential equations of these curves, we simply replace t by τ in (4.5) and dt by $ds = (dx^2 + dy^2 + dz^2)^{\frac{1}{2}}$. Then we get

$$(x + iy)(ds - dz) = (\tau - z + i\epsilon)(dx + i\,dy).$$

The solutions of this equation are given (apart from the spurious $x + iy = \text{const}$) by

$$\left.\begin{aligned} x^2 + y^2 + (z - \tau)^2 - 2\epsilon(x \sin\varphi \\ + y\cos\varphi)\tan\theta = \epsilon^2, \\ z - \tau = (x\cos\varphi - y\sin\varphi)\tan\theta, \end{aligned}\right\} \quad (4.8)$$

where θ and φ are constants defining the different curves. These curves are evidently circles, being intersections of spheres with planes. They twist around one another in such a way that every pair of circles is linked. The twisting has a *positive* screw sense if $\epsilon > 0$, i.e., if R_α is right-handed.[21] They lie on the set of coaxial tori[22] obtained by eliminating φ between the two equations. [These tori are the rotations about the z axis of a system of coaxial circles in the (x, z) plane.]

From the point of view of the completed space–time M, we should regard the hyperplane $t = \tau$ as being completed (conformally) by a point (namely I) at infinity. It then becomes topologically a three-dimensional sphere S^3 (of which the hyperplane $t = \tau$ may be regarded as the stereographic projection). The vector field on S^3 is everywhere non-singular and nowhere vanishing. The circles constitute an example of what is known as the Hopf fibring of S^3 (one example being "Clifford parallels" on S^3).

Note that all the circles in the hyperplane thread through the particular (smallest) circle of radius $|\epsilon|$ and centre $z = \tau$, $x = y = 0$ given when $\theta = 0$. If ϵ is small, this circle describes, as τ increases, a path approximating that of a null line. If ϵ is zero, the path is exactly the null line $z = t$, $x = y = 0$. For small ϵ, we may think of the lines of the Robinson congruence as defining an approximate null line, but the lines twist around one another and never quite meet. The twisting has a positive or negative screw sence according as R_α is right- or left-handed. [In the limit $\epsilon \to 0$, the circles (4.8) all touch the z axis at $z = \tau$. The tangents to these circles are orthogonal to the spheres touching the (x, y) plane at $z = \tau$, these spheres being the intersections of $z = \tau$ with the null cones with vertices on $z = t$, $x = y = 0$. The lines of the congruence are then just the generators of these null cones in this case, as expected.]

V. THE ASSOCIATED SPINOR FIELD OF A TWISTOR

We saw above that the tangent vectors to the lines of the Robinson congruence constituted a field of null vectors which was regular and nonvanishing everywhere on M. The general Robinson congruence R can be written compactly in spinor terms in a way which exhibits this (and other) facts very simply. This leads to an interpretation of twistors in terms of certain spinor fields.

Put $R_\alpha = \bar{L}_\alpha$, so we can use the notation of Secs. II and III. The condition $X^\alpha \bar{L}_\alpha = 0$, for X to belong to the congruence $\bar{L}$ is then, by (3.7), (3.2),

$$0 = \xi^A \bar{\mu}_A + \eta_{A'} \bar{\lambda}^{A'} = \xi^A(\bar{\mu}_A - i x_{AA'} \bar{\lambda}^{A'}),$$

from which it follows that ξ_A is proportional to the expression in the bracket. Since X defines X^α only up to proportionality, we are at liberty to choose the scale factor for ξ^A so that

$$\xi_A = \bar{\mu}_A - i x_{AA'} \bar{\lambda}^{A'}. \quad (5.1)$$

We can think of (5.1) as defining a *spinor field*, since ξ_A is a function of x_i, the associated null vectors $v^i \leftrightarrow \xi^{J'} \bar{\xi}^{J'}$ being tangents to the lines of

[21] The screw sense arising here depends, of course, on the "handedness" of the choice of matrices in (2.4).

[22] I am grateful to J. Terryl and J. E. Reeve for this observation concerning the Hopf fibring.

the congruence $\bar{L}$. Since $\lambda^A x_{AA'}\bar{\lambda}^{A'}$ is real, it follows that

$$\mathrm{Re}\,(\lambda^A\xi_A) = \mathrm{Re}\,(\lambda^A\bar{\mu}_A) = \tfrac{1}{2}L^a\bar{L}_a, \qquad (5.2)$$

so that ξ_A cannot vanish unless L^a is null. The field of null directions defined by (5.1), i.e., by the Robinson congruence $\bar{L}$, is thus *well defined and regular* throughout $M\{I\}$. The regularity at infinity also follows from the conformal transformation properties of (5.1) as we are to see shortly.

Now write

$$\nabla_{BB'} \equiv \partial/\partial x^{BB'}, \qquad (5.3)$$

so that $\nabla_{BB'}x_{AA'} = \epsilon_{AB}\epsilon_{A'B'}$. Then we have, from (5.1),

$$\nabla_{BB'}\xi_A = i\epsilon_{BA}\bar{\lambda}_{B'}. \qquad (5.4)$$

Thus, the field ξ_A satisfies the differential equation

$$\nabla^{B'}_{(B}\xi_{A)} = 0. \qquad (5.5)$$

Conversely, from this equation we can reconstruct the form of the expression (5.1). For (5.5) implies that for *some* spinor $i\bar{\lambda}_{B'}$, the derivative $\nabla_{BB'}\xi_A$ has the form given in (5.4). Furthermore, this $\bar{\lambda}_{B'}$ must be constant because $\nabla^{C'}_{C}\nabla^{B'}_{B}\xi_A = 0$ [as follows from (5.5), since skew-symmetry in B, A *and* in C, A implies that the expression vanishes]. Integration of (5.4) leads straight to (5.1), where $\bar{\mu}_A$ is another constant.

We now have the important result that *there is a one-to-one relation between twistors of valence* $[^0_1]$ *and spinor fields satisfying* (5.5). This gives us a more complete representation of a twistor than just as a Robinson congruence, since the *factor of proportionality* of L^a is now also represented. Note that $L^a\bar{L}_a$ is also expressible simply in terms of ξ_A as

$$L^a\bar{L}_a = \mathrm{Im}\,(\bar{\xi}^{B'}\nabla^A_{B'}\xi_A). \qquad (5.6)$$

But since we are interested in twistors in relation to the *whole* of M, rather than just $M\{I\}$, we should also ask how the field ξ_A is to be defined at infinity. Essentially all that is required here is to verify that (5.5) is invariant under conformal transformation [e.g., under inversions, since then I becomes a finite point,[7] cf., (7.18)]. This invariance is achieved[23] by specifying that ξ_A transform as a *conformal density of weight* $\frac{1}{2}$. Then the field equation (5.5) is also satisfied at infinity, in the sense that it holds with respect to some metric which is regular on the null cone of I. This, in particular, defines the Robinson congruence $\bar{L}$ regularly at infinity. (In fact, there is one line of $\bar{L}$ totally at infinity, namely that corresponding to the direction of λ^A.) The quantity (5.6) is also invariant[23] under conformal transformation (ξ_A of weight $\frac{1}{2}$), so we can refer to $L^a\bar{L}_a$ as the *conformal invariant* of the spinor field ξ_A or of the twistor L^a.

Some subtleties remain, however, concerning the extension of the field ξ_A across infinity. The possibility of defining two-valued (spinor) fields in $M\{I\}$ is well known. But here we require a spinor field defined *globally* on M. The multiple connectedness of M is such as to cause a slight difficulty in this respect. But in any case, we can think, instead, in terms of $\xi_A\xi_B$ or the associated bivector [cf., (2.7)], since these are not two-valued. Then we might expect to find a representation of a twistor up to sign. However, it turns out [cf., (7.25)] that twistors of odd total valence are *four*-valued under the conformal group. Thus, any such conformally invariant representation of a twistor (of odd valence), cannot distinguish the twistor from i times the twistor. (This is analogous to the fact that spinors of odd valence cannot be completely represented in terms of tensors, in a Lorentz covariant way, whereby the spinors are distinguished from their negatives.) In the present case, when we extend the field ξ^A across infinity, we find that the field on the two sides of the null cone of I cannot be exactly matched, but ξ^A on one side must be matched with $i\xi^A$ (or $-i\xi^A$) on the other side. [We can see this explicitly in terms of the inversion of (7.18), for which the conformal factor is $\Omega = x^l x_l/2a^2$. The null cone of I is transformed from the null cone, $\Omega = 0$, of the origin, across which Ω changes sign. The transformed ξ^A field picks up the factor $\Omega^{-\frac{1}{2}}$.] Thus, the spinor field (5.1), when defined over the *whole* of M, must be thought of as *four-valued*; if ξ_A is one value, then $i\xi_A$, $-\xi_A$, $-i\xi_A$ are the other three.

Since spinors can be represented geometrically (up to sign), this implies a geometrical realization of any twistor of valence $[^0_1]$ up to a multiple of 1, i, -1, or $-i$. The phase of ξ_A defines, as we have seen [cf., (2.7)], a half-plane element tangent to the light cone; $-\xi_A$ defines the same half-plane element; but $i\xi_A$ and $-i\xi_A$ define the opposite half-plane element. Thus, a complete unoriented null plane element is defined at each point by the phase of $\bar{L}_a$. As we follow one of the lines X belonging to the Robinson congruence $\bar{L}$, we find that this plane element rotates about X in the opposite direction from the neighboring lines of the congruence and twice as fast. Owing to the way connections are

[23] See Ref. 7, Eqs. (10.1), (10.7), and (10.8) for the relevant formulas.

made at infinity, this results in the aggregate of plane elements at points of X, effectively constituting a Möbius strip, which explains why the plane elements cannot be consistently oriented along X.

To see how these plane elements rotate about X, consider $(\xi^B\bar{\xi}^{B'}\nabla_{BB'})\xi_A = i\bar{\xi}^{B'}\bar{\lambda}_{B'}\xi_A$ [see (5.4)]. The real part of $2i\bar{\xi}^{B'}\bar{\lambda}_{B'}$ measures the rate of extension of the vector represented by ξ_A and the imaginary part measures the rate of rotation of the half-plane element about the null direction of ξ_A. By (5.2),

$$\text{Im}\,(2i\bar{\xi}^{B'}\bar{\lambda}_{B'}) = -2\,\text{Re}\,(\xi_B\lambda^B) = -L^a\bar{L}_a,$$

whence the half-plane element rotates in a negative or positive sense according as the twistor L^a is right or left handed. (It might appear that the rate of rotation is constant here but this is misleading since the scaling is different at different points of X.) Note that if L^a is null, there is no rotation of the half-plane element. We still get a Möbius strip, however, since a half-twist always occurs, in effect, at infinity.

To see how the neighboring lines of the congruence rotate about X, consider

$$\xi^A\bar{\xi}^{B'}\nabla_{BB'}\bar{\xi}_A = -i\bar{\xi}^{B'}\bar{\lambda}_{B'}\xi_B. \tag{5.7}$$

Here, the real part of $-i\bar{\xi}^{B'}\bar{\lambda}_{B'}$ measures the convergence of the null lines and the imaginary part measures their rotation about one another.[24] The rotation is in the opposite sense from that of the half-plane elements and half as fast.

The expansion of the null lines and the change in the magnitude of ξ_A are not conformally covariant concepts, whereas the rotations are. Another conformally covariant concept for a congruence of null lines is the *shear* of these lines.[11,24] In the present case we have

$$\xi^B\xi^A\nabla_{BB'}\xi_A = 0, \tag{5.8}$$

which states that the shear *vanishes*.

VI. THE COMPLEX PROJECTIVE SPACE C

We have seen that the null lines in M form an ∞^5 system which can be extended to an ∞^6 system by including "complexified" null lines (the latter being representable in terms of certain related structures called Robinson congruences). The members of this ∞^6 system can be given complex projective coordinates $(L^a) = (L^0, L^1, L^2, L^3)$. That is, it is just the three complex ratios $L^0 : L^1 : L^2 : L^3$ which are significant, the only restriction on $L^0, \cdots, L^3$ being that they must not *all* vanish. This ∞^6 system we may think of as constituting a *three-dimensional complex projective space* which we denote by the letter C. The "points" of C are just the "complexified" null lines (and the null lines) of M.

In fact, we have two alternative pictures of any given situation, namely the one in terms of M and the one in terms of C. For example, we may think of an object L, with projective coordinates (L^a) *either* as, say, a "complexified" null line of the real space–time M (M picture) *or* simply as a point in a certain projective three space (C picture). These are just two different ways of *visualizing* what is the same physical situation in each case. In order that the two pictures be completely equivalent, however, we need to be able to interpret, in C, the condition of *"reality"* of a null line in M, of *incidence* between null lines in M, and finally of *points* in M. In effect, this requires that the conjugation relation $L^a \leftrightarrow \bar{L}_a$ should have a meaning with regard to the C picture. Now, we have seen that a twistor L^a (valence $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$) refers to a point L of C; a twistor R_a (valence $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$) therefore refers to the dual concept of a *plane* R in C, namely the plane of all points X for which $X^aR_a = 0$. (This is, of course, a plane in the *complex* sense. As a *real* manifold it is a four-dimensional subset of the six-real-dimensional manifold C.) The conjugation relation $L^a \leftrightarrow \bar{L}_a$ therefore describes a point $\leftrightarrow$ plane correspondence in C, which we may refer to as a *Hermitian correlation* of signature $(++--)$. The signature here refers to the Hermitian form

$$X^a\bar{X}_a = X^0\overline{X^2} + X^1\overline{X^3} + X^2\overline{X^0} + X^3\overline{X^1}. \tag{6.1}$$

We regard this Hermitian correlation as being an *intrinsic* part of the geometric structure of C.

The (real) null lines in M are the points of the five-real-dimensional subset N (with topology $S^3 \times S^2$) of C defined by the equation $X^a\bar{X}_a = 0$. Thus, N is a hypersurface if we regard C as a real six-dimensional manifold, but it is not a hypersurface in the sense of the *complex* structure of C. We refer to the subset of C for which $X^a\bar{X}_a > 0$ holds as C^+ and the part for which $X^a\bar{X}_a < 0$ as C^-. The two sets C^+ and C^- are then disconnected from one another and have N as their common boundary (Fig. 4). If L is any point of C, we may regard the plane $\bar{L}$ as the *polar plane* of L with respect to N, since "polarizing" $X^a\bar{X}_a$ with L^a yields $X^a\bar{L}_a = 0$ (or $L^a\bar{X}_a = 0$), the equation of the plane $\bar{L}$. The Robinson congruence associated with L is now the intersection (topology S^3) of the plane $\bar{L}$ with N. (A slight inconsistency of notation arises here in that

[24] P. Jordan, J. Ehlers, and R. Sachs, Akad. Wiss. Lit. Mainz, no. 1 (Mainz 2) (1961); E. T. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962).

it is the Robinson congruence $\bar{L} \cap N$ that had been previously labeled $\bar{L}$, rather than the entire plane. The Robinson congruence includes only "real" null lines of M, by definition. In cases where there is possibility of confusion, we refer to "the Robinson congruence $\bar{L}$" or "the plane $\bar{L}$" as the case may be, but in general $\bar{L}$ refers to an entire plane in C.) When L lies on N, the plane $\bar{L}$ can be thought of as the *complex tangent plane to N at L*. This is also just the case when *L lies on its polar plane.*

Now, if we wish to represent a *point of M* in terms of the C picture, we may do this using incidence properties of null lines in M. Any point P in M can be uniquely represented by an ∞^2 system of null lines in M, namely by the generators of the null cone of P. Let K and L be two null lines in M through P. The generators of the null cone of P are then the null lines common to both $\bar{K}$ and $\bar{L}$ (i.e., the generators must meet both K and L). In the C picture this is an ∞^2 system[25] of lines on N which lie on the intersection of the two planes $\bar{K}$ and $\bar{L}$. This intersection is simply a complex projective (straight) line in C. Thus, we have the result that any point of M is represented, in the C picture, by a complex projective line which lies entirely on N. (A complex projective line is topologically a sphere S^2, in terms of its real structure. This agrees with the topology of the set of null lines through P, i.e., of the null directions at P.) We do not distinguish, notationally, here, between the point P in M and the system of null lines in M through P. Then we are at liberty to denote also by P the *line* in the C picture which represents this *point* P in the M picture.

Conversely, *any* line P in the C picture, which lies entirely on N, represents some point P in the M picture. (In the terminology of projective geometry, in the C picture, "line" always implies "complex projective straight line.") To see this, consider the C picture and let the line P lie entirely on N. Let K and L be two points on P. Then we have $K^a\bar{K}_a = 0$, $L^a\bar{L}_a = 0$, and, more generally $(K^a + \beta L^a)(\bar{K}_a + \bar{\beta}\bar{L}_a) = 0$ for all complex β (the general point on the line P having a twistor of the form $K^a + \beta L^a$). Hence $L^a\bar{K}_a = 0$. Thus, in the M picture, the null lines L and K must intersect. This holds for *any* pair of null lines belonging to the ∞^2 system represented by P. These null lines must therefore all meet in a *point* and, in fact, must be the generators of the null cone of this point. We thus label this point P, and the situation is as before.

[25] A symbol ∞^r indicates r dimensionality in the *real* sense.

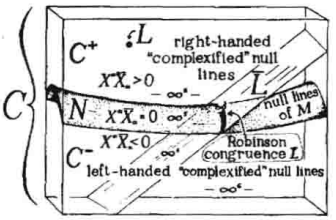


FIG. 4. The C picture.

Note that, in the C picture, the condition that a point L lies on a line P, both L and P lying on N, is interpreted in the M picture as the condition that the null line L passes through the point P. We have, in fact, a kind of duality correspondence between M and N. To sum up, we have the following correspondence between the M picture and the subset N in the C picture:

There is a one-to-one relation between the null lines in M and the points in N. (6.2)

There is a one-to-one relation between the points in M and the complex lines in N. (6.3)

The condition for a point to lie on a null line in M is interpreted, in N, as the condition for the corresponding line to pass through the corresponding point. (6.4)

From (6.4), we see that the condition that two points in M have a null separation is simply the condition that their corresponding lines in N should intersect; the condition that two null lines in M should meet is the condition that the join of the corresponding points in N should lie entirely in N.

We may ask how we should interpret the lines of C which do *not* lie entirely on N. Lines in a projective three space form a four-dimensional system, so that in terms of real dimensions these lines describe an ∞^8 system of objects in M. This suggests that a general line in C represents a *complexified* point in M, where now the straightforward doubling of dimensions suggests that these are complexified points in the usual sense, i.e., their position vectors are allowed to have imaginary parts. To see that this is indeed a consistent interpretation, consider a general line in the C picture as a join of two general points L, X of C. Using the notation of Secs. II and III, we can form the expression (3.10). We may regard this *complex vector* p_i as the position vector, the M picture, of the complex point of intersection of the two null lines ("complexified" or otherwise) in M, defined by L and X. (Note that in this sense *any*

two distinct null lines in M have a complex point in common.)

Any line P of C falls into one of six classes. If q^i is the *imaginary part* of the position vector p^i for the corresponding complex point P in M, then [from (3.10)]

$$P \subset C^+ \Leftrightarrow q^i \text{ timelike, future-pointing;} \tag{6.5}$$

$$P \subset C^+ \cup N,\ P \not\subset N,\ P \not\subset C^+ \Leftrightarrow q^i \text{ null, future-pointing;} \tag{6.6}$$

$$P \text{ intersects all of } C^-,\ N,\ C^+ \Leftrightarrow q^i \text{ spacelike, nonzero;} \tag{6.7}$$

$$P \subset C^- \cup N,\ P \not\subset N,\ P \not\subset C^- \Leftrightarrow q^i \text{ null, past-pointing;} \tag{6.8}$$

$$P \subset C^- \Leftrightarrow q^i \text{ timelike, past-pointing;} \tag{6.9}$$

$$P \subset N \Leftrightarrow q^i = 0. \tag{6.10}$$

In cases (6.6) and (6.8), the line P in C *touches* M. The set (6.5) (in the M picture) is sometimes referred to as a "future-tube."[10]

There is a representation of lines in projective three-space, well known to geometers, called the Klein (or Grassmann) representation.[26] This is a four- (complex-) dimensional quadric, the points of which correspond to the lines of complex projective three-space. Thus, in our case, we may regard the fully complexified version M^* of M as the Klein representation of the lines in C. We noted in Sec. III that M itself was essentially a real $(+\,+\,-\,-\,-\,-)$ quadric fourfold. M^* is the complexified version of this quadric. There are two systems of planes on M^*, called α planes and β planes. The α planes correspond to points in C and the β planes to planes in C. We may regard M as a submanifold of M^*. The α planes which meet M intersect M in the null lines of M. Similarly with the β planes. Thus, when properly complexified, the null lines in M become, in effect, pairs of complex projective planes, namely one α plane and one β plane for each fully complexified null line. The "complexification" of the null lines achieved here by the use of twistors amounts to selecting only the α planes to represent the null lines.

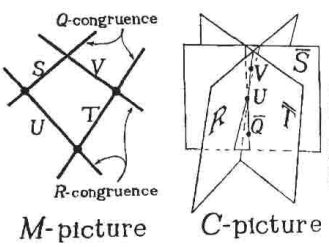


FIG. 5. The condition for $\bar{Q}^\alpha R_\alpha = 0$ is that, in the M picture, V and S should meet, and that in the C picture, $V \in \bar{S}$, whence $\bar{Q} \in R$.

Much of the earlier discussion could have been carried out in terms of the manifold M^*. The emphasis here, however, has been to try and represent twistors as far as possible in terms of "real" structures in the space–time M. One final condition remains to be represented, however, in order that the geometry of C can be completely realized in terms of M. This is that we must be able to interpret the vanishing of a twistor scalar product in geometrical terms[27] in M.

Consider the condition

$$\bar{Q}^\alpha R_\alpha = 0. \tag{6.11}$$

If $\bar{Q}$ and $\bar{R}$ are both null lines in M, we have seen that this is the condition for the lines to intersect. If $\bar{Q}$ is a null line and R a Robinson congruence, then (6.11) is simply the condition that $\bar{Q}$ should belong to the congruence R. It remains to consider the case when *both* Q and R are Robinson congruences in M. Let S and T be two null lines belonging to the Q congruence and let U and V be null lines belonging to the R congruence. Suppose the three pairs of lines (S, U), (U, T), (T, V) intersect. *Then a necessary and sufficient condition for* (6.11) *to hold is that the pair* (V, S) *also intersect* (Fig. 5). To see this, consider the points V, U, $\bar{Q}$ and the planes $\bar{S}$, $\bar{T}$, R in the C picture. We have $S \in Q$, $T \in Q$, whence $\bar{Q} \in \bar{S}$, $\bar{Q} \in \bar{T}$. Also $U \in R$, $V \in R$ and $U \in \bar{S}$, $U \in \bar{T}$, $V \in \bar{T}$. Thus, the line UV must be the intersection of the planes R and $\bar{T}$ (Fig. 5). If (6.11) holds, then Q also lies on this line, whence $V \in \bar{S}$. Conversely if $V \in \bar{S}$ then $\bar{Q} \in R$. Thus, in the M picture, the condition for (6.11) to hold is that the null lines V and S should meet.

This establishes the geometrical equivalence of the M picture with the C picture, whereby the conformal geometry of Minkowski space is completely represented in twistor terms. The *metric* geometry of Minkowski space, on the other hand, requires the introduction of a fixed metric twistor. This is done in Sec. X.

[26] See any standard work on classical algebraic geometry, for example, J. A. Todd, *Projective and Analytical Geometry* (I. Pitman, London, 1947); J. G. Semple and L. Roth, *Algebraic Geometry* (Clarendon Press, Oxford, England, 1949).

[27] Strictly, we should also show that the concept of a Robinson congruence is "geometrical" in M. An explicit construction in terms of incidence of null lines is given in Sec. IX, but in any case, the geometric (and conformally invariant) nature of a Robinson congruence is already implied by Sec. IV and V.

VII. TWISTOR TRANSFORMATIONS

Although the transformation properties of twistors have not been given explicitly as yet, they are implicit in the preceding discussion. Any continuous transformation of M into itself, which preserves its conformal structure (and, hence, the null cone and null line structure of M) must, by (6.4), correspond to a continuous transformation of N into itself which preserves its linearity structure. Furthermore, since the concept of a Robinson congruence is conformally invariant (cf., Sec. V), this transformation of N extends uniquely to a transformation of the whole space C. We have seen that the orthogonality relation (6.11) between points and planes in C can be stated in terms of (conformally invariant) incidence properties in M. Hence, collinearity of points in C has a conformally invariant interpretation in M. It follows that the conformal transformations of M are represented, in the C picture, by continuous point transformations of C, which preserve its linearity structure, and for which the submanifold N is invariant.

If we restrict ourselves to conformal transformations of M which are continuous with the identity, then the corresponding transformations of C is also continuous with the identity. Such transformations, preserving the linearity structure of C, must therefore be *projective point transformations*[26] of C (in the complex sense). If, on the other hand, we allow transformations of M which involve space or time reflections, then we must also consider *anti-projective* transformations of C, that is, transformations which combine a complex conjugation operation with a projective transformation (so that cross ratios become complex conjugated). We may also consider transformations of C which are not point transformations but *correlations*[26] (duality correspondences) in which points of C are mapped to planes, and planes to points. These would not strictly correspond to transformations of M, because of the way C has been constructed here, but it is convenient to think of them as representing transformations of the *complexified* version M^* of M. The *natural* (analytic) extension to M^* of a transformation defined on M sometimes turns out to correspond to a correlation in C.

The basic operation $\mathcal{C}$ of *complex conjugation* in M^*, which leaves the real space M invariant, corresponds, in the C picture, simply to the *Hermitian correlation* given by $X^\alpha \leftrightarrow \bar{X}_\alpha$. A transformation of M equivalent to a *space reflection* corresponds, in the C picture, to a point transformation of C, which is antiprojective and interchanges C^- and C^+. (The twist orientation of Robinson congruences is reversed under space reflection.) However, if we wish to extend this to a space reflection $\mathcal{P}$ of the whole of M^* we must, since $\mathcal{P}$ is analytic, remove the antiprojective nature of the transformation of C and consider, instead, a *correlation* which sends points X of $C^+[C^-]$ into planes $\bar{Y}$ for which Y is in $C^-[C^+]$. The antiprojective point transformation just considered, which interchanges C^+ and C^- would then correspond to $\mathcal{CP}$. Similarly, a point transformation of C which represents a *time reflection* of M is antiprojective, here transforming each of C^-, C^+ into itself. It therefore really represents $\mathcal{C}\mathfrak{I}$, where $\mathfrak{I}$ describes the (analytic) extension to M^* of this time reflection of M. The transformation of C representing $\mathfrak{I}$ is a *correlation* sending each point X of $C^+[C^-]$ into a plane $\bar{Y}$ with $Y \in C^+[C^-]$. The transformation $\mathcal{P}\mathfrak{I}$, being the product of two projective correlations, is a projective point transformation of C. It interchanges C^- with C^+, and is not continuous with the identity unless we widen the group of transformations of M^* to include *complex* conformal transformations. (These would not be point transformations of M, but we may view them as transformations on the line systems in M which transform Robinson congruences into one another. They are represented, in the C picture, as projective transformations of C which do not preserve N.) The transformation $\mathcal{P}\mathcal{C}\mathfrak{I}$ is an antiprojective correlation in C.

Let us examine more explicitly the effect of an allowable transformation on a general twistor $A_{\rho\cdots\tau}^{\alpha\beta\cdots\delta}$ of valence $[{}^r_s]$. (The indices $\alpha, \beta, \cdots, \delta$ are r in number and $\rho, \cdots, \tau$ are s in number, each ranging over four values 0, 1, 2, 3.) We may define a general twistor, in terms of twistors of valence $[{}^1_0]$ and $[{}^0_1]$, in any of the standard ways that "tensors" may be built up from "vectors," e.g., as linear combinations of outer products, or in terms of multilinear mappings of one-index twistors into the scalars. Alternatively, we may simply use the transformation properties to define a twistor $A_{\rho\cdots\tau}^{\alpha\cdots\delta}$ of valence $[{}^r_s]$ (considering for the moment only transformations continuous with the identity):

$$\bar{A}_{\rho\cdots\tau}^{\alpha\beta\cdots\delta} = A_{\varphi\cdots\psi}^{\kappa\lambda\cdots\nu} l_\kappa^\alpha l_\lambda^\beta \cdots l_\nu^\delta T_\rho^\varphi \cdots T_\tau^\psi. \tag{7.1}$$

The matrices (l_β^α), (T_β^α) are inverses[28] of each other

$$l_\beta^\alpha T_\gamma^\beta = \delta_\gamma^\alpha = T_\beta^\alpha l_\gamma^\beta, \tag{7.2}$$

[28] In fact, the transformations of C would be the same if we specified only that $(T_\beta{}^\alpha)$ be *proportional* to the inverse of $(l_\beta{}^\alpha)$ since the X^α are *projective* coordinates for C. However, the stronger requirement (7.2) is adopted here since the factor of proportionality of a twistor is required when the more complete representation in accordance with Sec. V is used.

where twistors X^α, Y_α, of respective valence $[^1_0]$, $[^0_1]$, transform as

$$\tilde{X}^\alpha = t^\alpha_\beta X^\beta, \qquad \tilde{Y}_\alpha = T^\beta_\alpha Y_\beta \tag{7.3}$$

under a conformal transformation of M continuous with the identity (i.e., a projective transformation of C preserving C^+). But since $\tilde{X}_\alpha$ is a twistor of valence $[^0_1]$, it follows that

$$T^\alpha_\beta = \bar{t}^\alpha_\beta, \quad \text{i.e.,} \quad t^\alpha_\beta \bar{t}^\beta_\gamma = \delta^\alpha_\gamma, \tag{7.4}$$

where the convention used earlier for twistor indices under complex conjugation [cf., (3.5)] is being employed here, namely that, under complex conjugation, upper and lower index positions are interchanged and also, the pairs 0, 1 and 2, 3 are interchanged, i.e.,[17]

$$\bar{t}^0_0 = t^2_2, \qquad \bar{t}^0_1 = t^3_2, \qquad \bar{t}^0_2 = t^0_2, \qquad \bar{t}^0_3 = t^1_2,$$
$$\bar{t}^1_0 = t^2_3, \cdots, \bar{t}^3_2 = t^0_1, \qquad \bar{t}^3_3 = t^1_1.$$

Condition (7.4) states that the matrix (t^α_β) is pseudo-unitary in the sense that the form (6.1) is left invariant, that is, in matrix notation,

$$(t^\alpha_\beta)^* \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} (t^\gamma_\delta) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \tag{7.5}$$

where the asterisk denotes Hermitian conjugate in the usual sense.[29]

Equation (7.4) ensures that the operation of complex conjugation, according to the rules just stated, is a twistor operation. That is, if $A^{\alpha\beta\cdots\delta}_{\rho\cdots\tau}$ is a twistor of valence $[^\cdot_\cdot]$, then $\tilde{A}^{\rho\cdots\tau}_{\alpha\beta\cdots\delta}$ is a twistor of valence $[^\cdot_\cdot]$ (where,[17] for example, $\tilde{A}^{3\cdots0}_{21\cdots3} = \overline{A^{03\cdots1}_{1\cdots2}}$, etc.). Clearly, the operations of addition, outer multiplication, contraction (between upper and lower indices), and index permutation (not mixing upper with lower indices) are also twistor operations. [That is to say, they commute with (7.1).] We might also consider complexified twistor transformations (7.1) for which (7.4) does not hold. These correspond to *complex* conformal transformations of M^*, not preserving M, and the relation between $A^{\alpha\cdots\delta}_{\rho\cdots\tau}$ and $\tilde{A}^{\rho\cdots\tau}_{\alpha\cdots\delta}$ would not be preserved. Such transformations will not be discussed here.

It will be appropriate to impose one final condition on the matrix (t^α_β), namely that it should be *unimodular*:

$$|t^\alpha_\beta| = 1. \tag{7.6}$$

This amounts to requiring that the Levi-Civita symbols $\epsilon_{\alpha\beta\gamma\delta}$, $\epsilon^{\alpha\beta\gamma\delta}$ [with fixed components $+1$, -1, 0 according as $(\alpha, \beta, \gamma, \delta)$ is an even permutation, an odd permutation, or no permutation of (0, 1, 2, 3)] shall be *twistors* rather than just "twistor densities." In fact, this condition is *necessary* if a strict interpretation of twistors according to the scheme of Sec. V is to be adhered to. The effect of (7.6) on the twistor algebra is only to enrich it slightly in that the operation of forming *duals* of skew-symmetric twistors is now allowable. Note also that

$$\epsilon_{\alpha\beta\gamma\delta} = \tilde{\epsilon}_{\alpha\beta\gamma\delta}, \qquad \epsilon^{\alpha\beta\gamma\delta} = \tilde{\epsilon}^{\alpha\beta\gamma\delta}, \tag{7.7}$$

since (2, 3, 0, 1) is an even permutation of (0, 1, 2, 3)

The transformations of the form (7.1) are not the only allowable twistor transformations since, as we have seen, the operations $\mathcal{C}$, $\mathcal{P}$, $\mathcal{T}$, $\mathcal{CP}$, $\mathcal{CT}$, and $\mathcal{PCT}$, give conformal transformations of M but they do not correspond to projective point transformations of C. On the other hand, $\mathcal{PT}$ *does* have the form (7.1), with (7.2) and (7.6) holding, but where (7.4) is replaced by

$$T^\alpha_\beta = -\bar{t}^\alpha_\beta; \qquad t^\alpha_\beta \bar{t}^\beta_\gamma = -\delta^\alpha_\gamma. \tag{7.8}$$

Thus, for $\mathcal{PT}$, we have $\tilde{\tilde{A}} \,\vdots\!:\, = -\tilde{A} \,\vdots\!:\,$ if the total valence of $A\,\vdots\!:\,$ is odd. (This, again, is assuming that we impose (7.2). Had we chosen $t^\alpha_\beta T^\beta_\gamma = -\delta^\alpha_\gamma$ instead, then $\tilde{\tilde{A}} \,\vdots\!:\, = \tilde{A} \,\vdots\!:\,$, *but* $\tilde{L}^\alpha \tilde{R}_\alpha = -L^\alpha R_\alpha$, etc.) The operation $\mathcal{C}$ is given by

$$A^{\alpha\beta\cdots\delta}_{\rho\cdots\tau} \to \tilde{A}^{\rho\cdots\tau}_{\alpha\beta\cdots\delta}, \tag{7.9}$$

while $\mathcal{PCT}$ is a combination of (7.9) with a transformation of the form (7.1) satisfying (7.8). The operations $\mathcal{P}$ and $\mathcal{T}$ belong to a class of twistor transformations given by

$$\tilde{A}^{\rho\cdots\tau}_{\alpha\beta\cdots\delta} = (\mp1)^s A^{\varkappa\lambda\cdots\nu}_{\varphi\cdots\psi} u_{\alpha\varkappa} u_{\beta\lambda} \cdots u_{\delta\nu} \bar{u}^{\rho\varphi} \cdots \bar{u}^{\tau\psi}, \tag{7.10}$$

where $(u_{\alpha\beta})$ and the transpose of $\mp(\bar{u}^{\alpha\beta})$ are unimodular inverse matrices:

$$u_{\beta\alpha}\bar{u}^{\beta\gamma} = \mp\delta^\gamma_\alpha; \qquad |u_{\alpha\beta}| = 1. \tag{7.11}$$

[29] The representation of twistors in terms of components given in this paper is perhaps the most convenient, but is by no means the only one possible. It amounts to insisting that the coordinate basis twistors $E^\alpha{}_{(0)}$, $E^\alpha{}_{(1)}$, $E^\alpha{}_{(2)}$, $E^\alpha{}_{(3)}$ are null and satisfy $E^\alpha{}_{(0)}\bar{E}_{(2)\alpha} = 1 = E^\alpha{}_{(1)}\bar{E}_{(3)\alpha}$, with all the other scalar products vanishing. Another possible choice would be to require $E^\alpha{}_{(a)}\bar{E}_{(b)\alpha} = (-1)^b\delta_{ab}$ (b not summed!) in which case twistor complex conjugation would take the form $\bar{L}_0 = \overline{L^0}$, $\bar{L}_1 = -\overline{L^1}$, $\bar{L}_2 = \overline{L^2}$, $\bar{L}_3 = -\overline{L^3}$, instead of (3.5), and the fixed matrix in (7.5) would become diag $(1, -1, 1, -1)$. (The simple connection with spinors (2.13) would be lost, however.) The only *essential* restriction on the way in which the twistors are represented is that the signature of the form $X^\alpha\bar{X}_\alpha$ must be $(++--)$ [cf., (6.1)].

The minus sign applies when a reversal of space orientation is involved and the plus sign when it is the time orientation that is reversed. Equation (7.11) implies that the matrix $(u_{\alpha\beta})$ is also pseudo-unitary in the sense of (7.4) and (7.5), or (7.8). These transformations give the projective correlations of C. The antiprojective point transformations of C such as $\mathfrak{CP}$ and $\mathfrak{CJ}$, are given by the combination of (7.10) with (7.9).

Let us now examine some twistor transformations in a little more detail. The *translation* given by

$$x^{i} = x^{i} + a^{i}, \qquad \xi^{A} = \xi^{A} \tag{7.12}$$

[using the notation of (3.1), (3.2)] results in

$$(\tilde{X}^{\alpha}) = (\xi^{A},\ \eta_{A'} - i\xi^{B}a_{BA'}),$$

whence[30]

$$(t^{\alpha}_{\beta}) = \begin{bmatrix} -\epsilon^{A}{}_{B} & 0 \\ -ia_{A'B} & \epsilon_{A'}{}^{B'} \end{bmatrix}. \tag{7.13}$$

The *Lorentz rotation* given, with $(b^{A}{}_{B})$ unimodular, by

$$x^{JJ'} = x^{KK'}b^{J}{}_{K}\bar{b}^{J'}{}_{K'}, \qquad \xi^{A} = \xi^{B}b^{A}{}_{B}, \tag{7.14}$$

results in

$$(\tilde{X}^{\alpha}) = (b^{A}{}_{B}\xi^{B},\ -\bar{b}_{A'}{}^{B'}\eta_{B'}),$$

whence

$$(t^{\alpha}_{\beta}) = \begin{bmatrix} b^{A}{}_{B} & 0 \\ 0 & -\bar{b}_{A'}{}^{B'} \end{bmatrix}. \tag{7.15}$$

Combining (7.14) with (7.12) and (7.15) with (7.13), we get the general restricted Poincaré transformation

$$x^{JJ'} = x^{KK'}b^{J}{}_{K}\bar{b}^{J'}{}_{K'} + a^{JJ'}, \qquad \xi^{A} = \xi^{B}b^{A}{}_{B} \tag{7.16}$$

[actually, "inhomogeneous $SL(2, C)$ transformation"[10]] represented by

$$(t^{\alpha}_{\beta}) = \begin{bmatrix} b^{A}{}_{B} & 0 \\ -ia_{CA'}b^{C}{}_{B} & -\bar{b}_{A'}{}^{B'} \end{bmatrix} \tag{7.17}$$

so that matrices of the form (7.17) give a (two-valued) representation of the Poincaré group.

The *inversion*

$$x^{i} = 2a^{2}x^{i}(x^{k}x_{k})^{-1} \tag{7.18}$$

$(a^{4} > 0)$ is given if we put $\xi^{A'} = a^{-1}\eta^{A'}$, $\tilde{\eta}_{A} = -a\xi_{A}$, i.e.,

$$\tilde{X}_{\alpha} = (-a\xi_{A},\ a^{-1}\eta^{A'})$$

since this agrees with (3.2) [and hence with (3.10)]. Thus, in this case

$$(u_{\alpha\beta}) = \begin{bmatrix} a\epsilon_{AB} & 0 \\ 0 & a^{-1}\epsilon^{A'B'} \end{bmatrix}. \tag{7.19}$$

Here a reversal in time orientation ($a^{2} > 0$) *or* space orientation ($a^{2} < 0$) is involved, so we get (7.10) rather than (7.1). The space–time reflection

$$x^{i} = -x^{i}, \qquad \xi^{A} = -\xi^{A} \tag{7.20}$$

is given if

$$\tilde{X}^{\alpha} = (-\xi^{A},\ \eta_{A'})$$

so that[30]

$$(t^{\alpha}_{\beta}) = \begin{bmatrix} \epsilon^{A}{}_{B} & 0 \\ 0 & \epsilon_{A'}{}^{B'} \end{bmatrix}. \tag{7.21}$$

For a space reflection or a time reflection, we must introduce a timelike vector v^{i} at the origin. For convenience we normalize v^{i} to have length $\sqrt{2}$, so that

$$x^{i} = \mp x^{i} \pm v^{i}v_{k}x^{k}; \qquad v_{k}v^{k} = 2. \tag{7.22}$$

The upper sign refers to space reflection and the lower sign to time reflection. In spinor terms we have

$$x^{JJ'} = \pm x^{KK'}v^{J}_{K}v^{J}_{K'}; \quad \xi^{A'} = \xi^{B}v^{A'}_{B}, \quad \tilde{\eta}_{A} = \pm\eta_{B'}v^{B'}_{A}$$

with $v^{B'}_{A}v^{C}_{B'} = -\delta^{C}_{A}$. Thus,

$$\tilde{X}_{\alpha} = (\pm v^{B'}_{A}\eta_{B'},\ v^{A'}_{B}\xi^{B}),$$

so that

$$(u_{\alpha\beta)} = \begin{bmatrix} 0 & \pm v_{A}{}^{B'} \\ v^{A'}{}_{B} & 0 \end{bmatrix}. \tag{7.23}$$

Note that $u_{\alpha\beta}$ is *symmetric* for a space reflection and *skew-symmetric* for a time reflection. Recall, also, that $u_{\alpha\beta}$ was skew-symmetric for the inversion (7.18) [cf., (7.19)]. The significance of this lies in the fact that there are two distinct kinds of projective correlation in C which are involutory (i.e., whose squares are the identity), namely polarity with respect to a quadric (corresponding to $u_{\alpha\beta}$ symmetric) and null-polarity with respect to a linear complex[26] ($u_{\alpha\beta}$ skew-symmetric). The null-polarity is distinguished by the fact that any point in C lies *on* the plane into which it is transformed, whence any null line in M meets the null line into which it is transformed. Thus, in addition to the cases of time reflection or inversion just considered, $u_{\alpha\beta}$ is skew-

[30] The staggering of the spinor indices is to indicate that, in each case, the left-hand index labels rows and right-hand index labels columns. Also, for notational consistency, $-\epsilon^{A}{}_{B}$, is used here instead of $\delta^{A}{}_{B}$, and $\epsilon_{A'}{}^{B'}$ instead of $\delta_{A'}{}^{B'}$, although they are all numerically equal!

symmetric in the case of a space reflection in a plane (that is, in a timelike hyperplane). The linear complex involved is, in each case, the system of lines in C which are left invariant by the correlation. Some of these lines lie in N corresponding to the points of M which are invariant under the transformation. These points constitute a *hypersphere* in M which is spacelike [as in (7.18) with $a^2 > 0$, or (7.22) with the lower sign] if a time reversal is involved, or timelike [e.g., (7.18) with $a^2 < 0$] if it is a change of spatial orientation that is involved. (A hyperplane is to be regarded as a case of a hypersphere in M.) Transformations for which $u_{\alpha\beta}$ is *symmetric* include, in addition to the space reflection in the origin (or, more correctly, in a timelike line), a reflection in a spacelike line which is accompanied by a time reflection. In these cases, the points of M left invariant by the transformation are represented by the generators which lie in N of the quadric (equation: $X^\alpha X^\beta u_{\alpha\beta} = 0$) defining the polarity in C. These invariant points in M constitute either a timelike circle (a timelike straight line being one case) or a pair of spacelike circles (a spacelike straight line together with a "circle" at infinity being one case).

The involutory projective *point* transformations of C are called harmonic perspectives.[26] They fall into two main classes, depending on whether the invariant points of C constitute two skew lines (when $t^\alpha_\alpha = 0$) or a point and a plane ($t^\alpha_\alpha = \pm 2, \pm 2i$). The *first* class is further subdivided according as the two skew lines both lie in N, both cross N, or lie one in C^+ and one in C^-. (No other cases are possible.) If the lines both lie on N, they correspond to two special invariant points of M. The spacelike 2-sphere in M of intersection of the null cones of these two special points consists also of invarient points. Examples of this type of transformation are the space–time reflection (7.20) in the origin O (whence O and I are the special invariant points and the invariant 2-sphere is at infinity) or a reflection in a plane together with a time reversal (in which case the plane is the invariant "2-sphere" and the special invariant points are both at infinity). If the two skew lines both cross N, we get a timelike 2-sphere of invariant points in M (e.g., a timelike 2-plane, for the case of an ordinary reflection in a line). If neither skew line meets N, there are *no* invariant points in M [e.g., the inversion (7.18) with $a^2 > 0$, followed by the time reflection of (7.22)]. The transformations of the *second* class are also interesting in that no points of M are left invariant. The plane R of invariant points of C meets N in a set of points representing a Robinson congruence. The lines of the Robinson congruence are all invariant under the transformation but not pointwise invariant. Noninvolutory transformations with this property also exist. These are intimately related to twistors and their *four-valuedness* and so will be described briefly next.

Choose a right-handed twistor R_α normalized so that

$$R_\alpha \bar{R}^\alpha = 1.$$

Define

$$t(\theta)^\alpha_\beta = e^{i\theta}(\delta^\alpha_\beta - \bar{R}^\alpha R_\beta) + e^{-3i\theta}\bar{R}^\alpha R_\beta \qquad (7.24)$$

for each real θ. Then $t(\theta)^\alpha_\beta$ satisfies (7.4) and (7.6), so it represents an allowable twistor transformation. Also

$$t(\theta)^\alpha_\beta t(\varphi)^\beta_\gamma = t(\theta + \varphi)^\alpha_\gamma$$

so that these transformations form a one-parameter subgroup of twistor transformations. We can represent any line L of the *Robinson congruence* R, in M, by a twistor L^α satisfying $L^\alpha R_\alpha = 0$ (and $L^\alpha \bar{L}_\alpha = 0$). Thus,

$$t(\theta)^\alpha_\beta L^\beta = e^{i\theta} L^\alpha$$

whence each line of R must be left invariant under the transformation. Furthermore, except for those values

$$\theta = \tfrac{1}{2}n\pi \qquad (n = \cdots, -2, -1, 0, 1, 2, \cdots)$$

for which $t(\theta)^\alpha_\beta$ is proportional to δ^α_β, there are *no points* of M invariant under the transformation. (This follows because, in the C picture, the lines which are left invariant are those which either pass through the point $\bar{R}$ or lie in the plane R. In neither case can these lines lie on N, since $\bar{R}$ is not on N and R does not touch N.)

Finally, observe that for $\theta = \frac{1}{2}\pi$ we have

$$t\left(\frac{\pi}{2}\right)^\alpha_\beta = i\,\delta^\alpha_\beta. \qquad (7.25)$$

This gives the identity transformation on C and therefore also the identity transformation on M. But it *multiplies every twistor of valence* $[^1_0]$ *by* i. Also (7.25) is continuous with the identity twistor transformation via (7.24) with $0 \leq \theta \leq \frac{1}{2}\pi$. Hence *twistors are essentially four-valued under conformal transformations of* M.

VIII. THE KERR THEOREM

Consider the problem of finding all null solutions of Maxwell's equations (i.e., with $F_{ij}F^{ij} = 0 =$

$F_{ij}F_{kl}\epsilon^{ijkl}$) in free space. Robinson[11] showed that the problem could be reduced to that of finding all the shear-free null geodetic congruences in the space–time [cf., (5.8)], since associated with any null field was a congruence of this type, while conversely, given such a congruence, all the corresponding solutions of Maxwell's equations could then be defined in terms of arbitrary complex functions. The result applies to curved space–times as well as flat. We are concerned here only with space–times which are (conformally) flat. With this restriction, Robinson's result applies also to null zero rest-mass fields of arbitrary spin ($>\frac{1}{2}$). (The result for spin 2 has a bearing on the construction of null solutions of Einstein's nonlinear field equations.[31])

The solution of the remainder of the problem for flat space–time, namely the construction of all the shear-free null congruences, is given by a remarkable theorem due to Kerr.[32] The theorem takes the following very natural form when stated in twistor terms.

> *A congruence of null lines in M is shear-free if and only if it is representable in C as the intersection of N with a complex analytic surface S in C (or as a limiting case of such an intersection).* (8.1)

"Complex analytic" means here analytic in the sense of the complex structure of C, where S is a "surface" in the complex sense (i.e., it has four real dimensions). Thus, (8.1) tells us that the shear-free condition [cf., (5.8)] is, effectively, a Cauchy–Riemann type of relation in the C picture. [The reason for the final parenthetic remark in (8.1) is to enable certain exceptional cases to be incorporated, for which the congruence does not form an analytic system even in the real sense. These exceptional cases appear to occur only when the rotation of the congruence also vanishes, cf., (5.7).]

Suppose, first, that a congruence of null lines in M is defined by an equation

$$\varphi(X^\alpha) \equiv \varphi(\xi^A, \eta_{A'}) = 0, \tag{8.2}$$

where φ is analytic and homogeneous in the four complex variables X^α so that (8.2) defines a complex analytic surface S in C. We see that the congruence must be shear free. Putting $\eta_{A'} = -i\xi^A x_{AA'}$, as in (3.2), Eq. (8.2) can be used to solve for ξ^A, up to proportionality, as a function of $x_{AA'}$, i.e., to give the direction of the line X as a function of a point on X with position vector x^i. Euler's condition on φ, for it to be homogeneous (degree n), gives

$$X^\alpha \frac{\partial\varphi}{\partial X^\alpha} \equiv \xi^A \frac{\partial\varphi}{\partial\xi^A} + \eta_{A'} \frac{\partial\varphi}{\partial\eta_{A'}} = n\varphi.$$

Hence, by (3.2) and (8.2),

$$\xi^A\left(\frac{\partial\varphi}{\partial\xi^A} - ix_{AA'}\frac{\partial\varphi}{\partial\eta_{A'}}\right) = 0,$$

whence

$$\frac{\partial\varphi}{\partial\xi^A} - ix_{AA'}\frac{\partial\varphi}{\partial\eta_{A'}} = \kappa\xi_A \tag{8.3}$$

for some κ. The derivative of (8.2) with respect to $x_{AA'}$ [cf., (5.3)] must also vanish

$$\begin{aligned} 0 = \nabla^{BB'}\varphi &= \frac{\partial\varphi}{\partial\xi^A}\nabla^{BB'}\xi^A + \frac{\partial\varphi}{\partial\eta_{A'}}\nabla^{BB'}\eta_{A'} \\ &= \frac{\partial\varphi}{\partial\xi^A}\nabla^{BB'}\xi^A \\ &\quad + \frac{\partial\varphi}{\partial\eta_{A'}}(-ix_{AA'}\nabla^{BB'}\xi^A - i\xi^A\nabla^{BB'}x_{AA'}) \\ &= \kappa\xi_A\nabla^{BB'}\xi^A - i\frac{\partial\varphi}{\partial\eta_{B'}}\xi^B, \end{aligned}$$

by (3.2) and (8.3). Assuming $\kappa \neq 0$, contraction with ξ_B gives $\xi_B\xi_A\nabla^{BB'}\xi^A = 0$, i.e., the condition (5.8) for the null directions defined by ξ^A to be tangent to *shear-free* null straight lines, as required. Finally, κ *cannot* vanish if (8.2) represents a genuine condition ξ^A, since the left-hand side of (8.3) is simply the derivative, with respect to ξ^A, of φ considered as a function of ξ^A and $x_{AA'}$.

The converse result that (8.2) represents essentially the most *general* shear-free congruence in M is somewhat less straightforward because of the existence of exceptional cases. However, if we assume that the congruence is *analytic in the real sense*—and any nonanalytic congruence (presumably necessarily rotation-free) can be approximated arbitrarily closely by analytic ones—then we can see from the form of (5.8) that the congruence is *determined* once the directions of the lines (i.e., of ξ^A) are known on any *spacelike 2-sphere* F in M. [If $\xi^0/\xi^1 = \zeta$, then (5.8) becomes $\partial\zeta/\partial x^{00'} = -\zeta\partial\zeta/\partial x^{10'}$, $\partial\zeta/\partial x^{11'} = -\zeta^{-1}\partial\zeta/\partial x^{01'}$. This defines the propagation of ζ

---

[31] I. Robinson and A. Trautman, Proc. Roy. Soc. (London) **A265**, 463 (1962).

[32] R. P. Kerr, private communication; cf. also R. P. Kerr, Phys. Rev. Letters **11**, 238 (1963); R. P. Kerr and A. Schild, in Proceedings of the American Mathematical Society Symposium, April 1964. According to Kerr's original construction, a general shear-free null geodetic congruence is defined by an analytic relation $\phi(\eta, x^{00'} + \eta x^{10'}, x^{01'} + \eta x^{11'}) = 0$, where $dx^{00'} + \eta dx^{10'} = 0 = dx^{01'} + \eta dx^{11'}$.

off the 2-plane $x^{00'} = x^{11'} = 0$. On this 2-plane—which from the point of view of M is a 2-sphere—ζ can be given arbitrarily as a function of the complex variable $x^{01'}$ and its complex conjugate $x^{10'}$.] The 2-sphere F is the intersection of two null cones (vertices A, B) in M and is therefore represented, in the C picture by the lines lying in N which meet two skew lines A, B in N. The *complexification* F^* of F is, in the C picture, the system of *all* lines in C meeting both A and B. Now (in the M picture), F is a *real environment*[10] for F^*, so that any complex analytic function on F extends uniquely to a complex analytic function on F^*. The null directions of the congruence, defined at points of F, are represented in the C picture as an ∞^2 system of points lying on the lines of the F system (describing a *real* analytic surface). These extend uniquely to an ∞^4 system of points S on the lines of the F^* system (describing a *complex* analytic surface). The intersection of S with N then defines the given shear-free congruence as required (since the congruence defined by S agrees with the given one on F).

The Kerr theorem provides a very convenient means of studying the structure of shear-free null congruences in M, in general. Only a few results are briefly indicated here. For example, the lines of a shear-free null congruence along which the *rotation vanishes* are represented in C by the points where lines of N touch S. We can also generally form the *reciprocal*[26] $\tilde{S}$ of S with respect to the Hermitian correlation defined by N (i.e., the envelope of polar planes with respect to N of points of S). Then $\tilde{S}$ defines a shear-free null congruence in M which is, in a sense, "reciprocal" to the original one. The congruences which are *everywhere rotation-free* are the ones which are *self-reciprocal* (although the individual lines of the congruence do not generally reciprocate to themselves). In this case, S is a *ruled surface*[26] in C, with ∞^1 of its generators lying in N. This ∞^1 system corresponds, in M, to a *curve* and the lines of the congruence are just the null lines meeting this curve.[33] This curve in M is *null* if the ruled surface in C is *developable*.

Of special interest are the *algebraic* shear-free null congruences.[34] In this case S is algebraic variety[26] and φ can be given as a homogeneous polynomial

$$\varphi(X^\alpha) \equiv S_{\alpha\beta\ldots\delta}X^\alpha X^\beta \cdots X^\delta = 0, \qquad (8.4)$$

where $S_{\alpha\beta\ldots\delta}$ is symmetric, of valence $[^0_n]$. In terms of ξ^A and $\eta_{A'} = -i\xi^A x_{AA'}$, (8.4) becomes

$$\xi^A\xi^B \cdots \xi^D\Phi_{AB\ldots D} = 0, \qquad (8.5)$$

where the spinor field $\Phi_{AB\ldots D}$ is defined by

$$\Phi_{AB\ldots D} = {}^0\sigma_{(AB\ldots D)} + {}^1\sigma_{A'(B\ldots D}x^{A'}_{A)} + \cdots + {}^n\sigma_{A'B'\ldots D'}x^{A'}_{(A}x^{B'}_{B}\cdots x^{D'}_{D)}, \qquad (8.6)$$

the ${}^r\sigma\ldots$'s being constants which are essentially the coefficients $S_{\alpha\beta\ldots\delta}$. (The round brackets denote symmetrization.) The $\Phi_{AB\ldots D}$ is symmetric, satisfies the "field equation"

$$\nabla^{E'}_{(E}\Phi_{AB\ldots D)} = 0 \qquad (8.7)$$

by virtue of (8.6) and has a "canonical decomposition"[7]

$$\Phi_{AB\ldots D} = {}^1\xi_{(A}{}^2\xi_B \cdots {}^n\xi_{D)}, \qquad (8.8)$$

where each ξ of (8.8) satisfies (8.5). Conversely, *every* solution of (8.7) has the form (8.6) and each resulting ξ of (8.8) [or (8.5)] defines (one branch of) the corresponding algebraic shear-free null congruence. This generalizes the results of Sec. V (for which $n = 1$) and gives us a representation of an arbitrary symmetric twistor of valence $[^0_n]$, in terms of a symmetric spinor field satisfying (8.7).

Robinson's construction of the general, null, zero rest-mass field from its associated shear-free congruence can also conveniently be represented in twistor terms: the field can be defined in terms of a complex function on S. However, such matters are not entered into here. The twister description of physical fields, generally, is left to a later paper.

IX. GEOMETRICAL APPLICATIONS OF TWISTORS

Many interesting geometrical properties arise from the interplay between the geometric structure of M^* and that of C. Some of these result simply from the fact that the former is the Klein representation of the latter and are, therefore, essentially classical results of algebraic geometry.[26] Others, however, take into account the reality structure of M and so have a more direct relevance to the structure of the physical world. The natural algebra of the C picture—namely twistor algebra—can be used to derive certain geometrical properties of M. A small selection is given here.

A linear space of dimension r in C can be represented by a *(simple) skew-symmetric twistor* of valence $[^{r+1}_0]$ or by its *dual* of valence $[^0_{3-r}]$, $(r =$

[33] The rotation-free, shear-free null congruences which are *not* analytic in the real sense emerge here simply as the system of null lines meeting a *nonanalytic* curve in M.

[34] Explicit shear-free null congruences of this type have been used to generate explicit solutions of Einstein's equations. For example, in Kerr's construction of the field of a rotating body, $\varphi(x^\alpha)$ is quadratic. For details, see Ref. 32.

0, 1, 2). The relation between general skew-symmetric twistors and their duals is

$$A_{\alpha\beta\gamma} = A^{\delta}\epsilon_{\alpha\beta\gamma\delta}, \qquad B_{\alpha\beta} = \tfrac{1}{2}B^{\gamma\delta}\epsilon_{\alpha\beta\gamma\delta}, \qquad C_{\alpha} = -\tfrac{1}{6}C^{\beta\gamma\delta}\epsilon_{\alpha\beta\gamma\delta}, \tag{9.1}$$

where conversely

$$A^{\alpha} = -\tfrac{1}{6}A_{\beta\gamma\delta}\epsilon^{\alpha\beta\gamma\delta}, \qquad B^{\alpha\beta} = \tfrac{1}{2}B_{\gamma\delta}\epsilon^{\alpha\beta\gamma\delta}, \qquad C^{\alpha\beta\gamma} = C_{\delta}\epsilon^{\alpha\beta\gamma\delta}. \tag{9.2}$$

Because of (7.7) and the symmetry between (9.1) and (9.2), the operations of complex conjugation and of forming the dual commute. The meaning of, for example, $\bar{B}^{\alpha\beta}$, given $B^{\alpha\beta}$, is therefore unambiguous. We call $B^{\alpha\beta}$ *real* if $B^{\alpha\beta} = \bar{B}^{\alpha\beta}$ (whence also $B_{\alpha\beta} = \bar{B}_{\alpha\beta}$).

A skew-symmetric twistor $P^{\alpha\beta}$ is *simple* if it is of the form

$$P^{\alpha\beta} = X^{\alpha}Y^{\beta} - Y^{\alpha}X^{\beta}, \tag{9.3}$$

whence $P_{\alpha\beta}$, $\bar{P}^{\alpha\beta}$, $\bar{P}_{\alpha\beta}$ are all also simple. ($C^{\alpha\beta\gamma}$ and $A_{\alpha\beta\gamma}$ are necessarily simple.) Equivalent alternative conditions for $P^{\alpha\beta}$ to be simple are

$$P^{\alpha\beta}P_{\alpha\gamma} = 0 \quad \text{or} \quad P^{\alpha\beta}P_{\alpha\beta} = 0. \tag{9.4}$$

As given by (9.3), $P^{\alpha\beta}$ represents, in the C picture, the *line* joining two points X and Y. In the M picture, $P^{\alpha\beta}$ represents a complex *point* P (i.e., point of M^*). Thus, we can use *simple skew-symmetric twistors* $P^{\alpha\beta}$ for *projective coordinates in* M^*.

Now, the Hermitian correlation defined by N effects the correspondence $P^{\alpha\beta} \leftrightarrow \bar{P}_{\alpha\beta}$. This is the complex conjugation operation in M^* [cf., (7.9)] so, in terms of our coordinates $P^{\alpha\beta}$, complex conjugation in M^* is defined by $P^{\alpha\beta} \leftrightarrow \bar{P}^{\alpha\beta}$, the points of M being given when $P^{\alpha\beta}$ is *real* (in the above sense). Corresponding to (6.5)–(6.10) we can also express, in terms of $P^{\alpha\beta}$, the space–time nature of the imaginary part q^i of the position vector of a complex point P of M^*:

$$P^{\alpha\beta}\bar{P}_{\alpha\beta} \gtreqless 0 \text{ according as } q^iq_i \gtreqless 0, \tag{9.5}$$

and if $q^iq_i \geq 0$, then for all Z_α

$$Z_\alpha\bar{Z}^\beta P^{\alpha\gamma}\bar{P}_{\beta\gamma}\begin{cases}\geq 0 & \text{if } q^i \text{ future-pointing,}\\ \leq 0 & \text{if } q^i \text{ past-pointing.}\end{cases} \tag{9.6}$$

[The point of intersection of the plane Z with the line P in C is represented by $Z_\alpha P^{\alpha\gamma}$ and, cf., (6.5), (6.6), (6.8), and (6.9).]

A complex point of M^* can be realized in terms of a *real* structure in M in various ways; for example, in terms of the involutory transformation (see Sec. VII) which is represented in C by the harmonic perspective whose axes are the lines P and $\bar{P}$ in C (assuming P and $\bar{P}$ are skew, i.e., $P_{\alpha\beta}\bar{P}^{\alpha\beta} \neq 0$); or in terms of linear systems of Robinson congruences. However, if q^i is *spacelike*, we have a more easily visualizable representation.[35] In this case, the lines P, $\bar{P}$ in C each meets N in an ∞^1 system of points. In the M picture, these become two ∞^1 systems of null lines. No two null lines belonging to the same system can intersect, but every null line of the P system meets every null line of the $\bar{P}$ system (since in the C picture, all the points of the line P are conjugate under the Hermitian correlation to all the points of the line $\bar{P}$.) We can refer to the P system of lines in M as a *null regulus* and the $\bar{P}$ system as its *complementary* null regulus.[26] Any two nonintersecting null lines in M belong to a unique null regulus; the null transversals of these two null lines generate the complementary null regulus. The null reguli in M thus geometrically represent the points in M^* whose position vectors have spacelike imaginary parts.

A *real* skew-symmetric twistor $B^{\alpha\beta}(=\bar{B}^{\alpha\beta})$, which is *not* simple, also has a direct interpretation in M. We can, in fact, normalize $B^{\alpha\beta}$ so that

$$B^{\alpha\beta}B_{\alpha\gamma} = \pm\delta^{\beta}_{\gamma}. \tag{9.7}$$

Then, if we put $B_{\alpha\beta} = u_{\alpha\beta}$ in (7.10) we get an involutory projective correlation, in C, of the type for which the invariant points in M constitute a hypersphere (cf., Sec. VII). Thus, $B^{\alpha\beta}$ represents a *hypersphere* in M which is spacelike or timelike according as the upper or lower sign occurs in (9.7). (In the limiting case when $B^{\alpha\beta}$ becomes simple, the hypersphere becomes a null cone with vertex B.)

We have seen that A^α represents a point A in C and that $\bar{A}_\alpha$ represents its "polar plane" with respect to N (cf., Sec. VI). A lies on N if and only if $A^\alpha\bar{A}_\alpha = 0$, which is also the condition for $\bar{A}$ to touch N. In terms of the dual twistor $A_{\alpha\beta\gamma}$, this condition is

$$\bar{A}^{\alpha\beta\gamma}A_{\alpha\beta\gamma} = 0. \tag{9.8}$$

Let X^α, Y^α, Z^α be three (nonzero) null twistors so

[35] When q^i is timelike we may represent P as a parallelism on M (with torsion; left-handed if $P\subset C^+$) which is closely related to Clifford parallelism on S^3. The sets of *null* directions which are to be regarded as parallel are those of the Robinson congruences represented by the points of the line P in C (i.e., by planes through $\bar{P}$). A transitive four-parameter group of (conformal) motions of M preserves this parallelism, namely that given by twistor transformations (7.1) for which the line P is left pointwise invariant. This group is readily seen to be the group of unitary (2×2) matices and leads to Uhlmann's representation (see Ref. 18) of the points of M in terms of such matrices.

FIG. 6. A geometrical theorem concerning null lines in M.

that, in the M picture, X, Y, and Z are three null lines. Put

$$\bar{A}^{\alpha\beta\gamma} = X^{[\alpha}Y^{\beta}Z^{\gamma]}. \tag{9.9}$$

(Square brackets denote skew-symmetrization.) In the M picture, A is generally a "complexified null line". But if (9.8) is satisfied, then A is a real null line in M which meets each of X, Y, and Z. ($\bar{A}_{\alpha}X^{\alpha} = 0$ follows from $\bar{A}^{[\alpha\beta\gamma}X^{\delta]} = 0$, etc.) Thus, *the condition for the three null lines X, Y, Z to have a common null transversal* is simply (9.9) substituted into (9.8), i.e.,

$$\{XY\}\{YZ\}\{ZX\} = 1, \tag{9.10}$$

where

$$\{XY\} = -(X^{\alpha}\bar{Y}_{\alpha})(Y^{\beta}\bar{X}_{\beta})^{-1} = \{YX\}^{-1}$$

etc. (since X^{α}, Y^{α}, and Z^{α} are null). We assume no two of X, Y, Z meet so that (9.10) is well defined. If X, Y, and Z have *two* common transversals, then (9.9) vanishes and X, Y, and Z belong to a null regulus.

It is possible to derive a host of geometrical theorems concerning incidence of null lines in M, from the condition (9.10). For example, if four null lines in M, of which no two meet, are such that there is a null transversal to each of three different selections of triplets of the lines, then there is also a null transversal to the remaining triplet. (The configuration is that depicted in Fig. 6—except for the case when there is a single common transversal to all four lines.) Some theorems of this general type can be generated by means of a diagrammatic notation: a graph made up of triangles whose *vertices* represent null lines in M, can be used. The triangles represent triples of lines with a common null transversal. Now, if any one of these triangles represents a circuit in the graph linearly dependent on the circuits given by other triangles, then, because of the form of (9.10), we have a geometrical theorem. For the case just considered (cf., Fig. 6), the graph would be that of a tetrahedron.[36] Any other polyhedron of triangles would also give rise to a theorem on null lines.

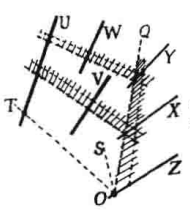


FIG. 7. Construction of the Robinson congruence containing three given null lines U, V, and W.

Finally, let us consider the *construction of a Robinson congruence* R containing three given null lines U, V, W in general position in M. (In the C picture, R is the plane of the three points U, V, W.) We wish to construct the line of R which passes through a general point O in M. Now, a Robinson congruence is characterized by the fact that, if any two null lines X, Y belong to the congruence, then so does every line of the null regulus containing X and Y. (In the C picture, the join of any two points in a plane also lies on the plane.) We have seen that the null regulus containing X and Y is simply the set of null transversals of any pair of null transversals of X and Y. Thus, if we can find X and Y belonging, respectively, to the null regulus containing U, V and to that containing U, W such that the null regulus containing X and Y also contains a line Z through O (see Fig. 7), the construction is complete. What is required, in fact, is to show that there is a null line Q through O which meets a line (namely X) of the U, V regulus and a distinct line (namely Y) of the U, W regulus; i.e., the triplets U, V, Q and U, W, Q must each have a null transversal. In fact, there is, in general, a unique such line Q through O. For, putting $Q^{\alpha} = S^{\alpha} + \zeta T^{\alpha}$, where T and S are null lines through O, with T meeting U, and substituting in (9.10), we get two simultaneous linear equations in ζ and $\tilde{\zeta}$. The condition for these equations to have a unique solution reduces to $\{UV\}\{VT\}\{TW\}\{WU\} \neq 1$ which, for general positions of O, is indeed satisfied.[37]

X. THE METRIC TWISTOR

If we wish to use twistors to describe the Minkowski metric structure of M (i.e., $M\{I\}$), rather than just its conformal structure, we may do this by introducing a *metric twistor* $I^{\alpha\beta}$ which represents the point I of M, according to the scheme of Sec. IX. Thus, $I^{\alpha\beta}$ is simple, skew-symmetric, and real (in the sense of Sec. IX):

$$I^{\alpha\beta}I_{\alpha\gamma} = 0, \tag{10.1}$$

$$I^{\alpha\beta} = \bar{I}^{\alpha\beta} \tag{10.2}$$

[36] Since a real null line in M defines both a point in C and a plane in C through this point, the configuration of Fig. 6 is represented in the C picture as a pair of mutually inscribed and circumscribed tetrahedra—a configuration familiar to geometers. We may note that the *full* complexification of a null line in M leads, in the C picture, strictly to a point in C *together* with a plane through it. This gives a five-complex-dimensional system as we would expect.

[37] It only fails if there is a circle through O meeting U, V, and W which lies on a null cone through U.

(with $I^{\alpha\beta} = -I^{\beta\alpha}$). The twistor transformations which leave $I^{\alpha\beta}$ invariant give the *Poincaré group*. (The transformations which leave $I^{\alpha\beta}$ invariant up to *proportionality* give the Poincaré group plus the *dilatations*. These would be just the conformal transformations of M which leave the point I—and its null cone—invariant.)

We may also treat *de Sitter space-time* in a very similar way. In this case there is a hypersphere at infinity rather than a null cone at infinity. As we have seen in Sec. IX, this can be described by a real, skew-symmetric twistor which is *not simple*, i.e., we just drop condition (10.1). Other conformally flat space-times can also be treated. For example, if we retain (10.1) but drop (10.2) [and replace (10.3) by its modulus], we can describe the Einstein static universe. The matter is not pursued further here, however.

We can generate Poincaré covariant operations simply by employing twistor algebra and admitting, as basic elements of the algebra (in addition to $\epsilon_{\alpha\beta\rho\sigma}$, $\epsilon^{\alpha\beta\rho\sigma}$, δ^{α}_{β}), the twistors $I^{\alpha\beta}$, $I_{\alpha\beta}$, satisfying (10.1), (10.2). We can represent points P, of $M^*\{I\}$ by simple skew-symmetric twistors $P^{\alpha\beta}$ normalized, for convenience, so that

$$P^{\alpha\beta}I_{\alpha\beta} = 2. \tag{10.3}$$

($P \in M\{I\}$ if $P^{\alpha\beta} = \bar{P}^{\alpha\beta}$.) Then the correspondence between P and $P^{\alpha\beta}$ is unique.[38] If $Q^{\alpha\beta}$ and $R^{\alpha\beta}$ similarly represent points of $M^*\{I\}$, then an example of a Poincaré covariant operation is

$$aP^{\alpha\beta} + bQ^{\alpha\beta} + cR^{\alpha\beta} - \tfrac{1}{2}\{bcQ^{\gamma\delta}R_{\gamma\delta} + caR^{\gamma\delta}P_{\gamma\delta} + abP^{\gamma\delta}Q_{\gamma\delta}\}I^{\alpha\beta}. \tag{10.4}$$

The significance of this particular expression is that if

$$a + b + c = 1 \tag{10.5}$$

then (10.4) represents the point in $M^*\{I\}$ whose position vector is $ap^i + bq^i + cr^i$, where p^i, q^i, r^i are the respective position vectors of P, Q, R. [This vector operation is clearly Poincaré covariant if (10.5) holds and represents a weighted mean.] Expression (10.4) generalizes to any number of points in an obvious way. Other less involved expressions than (10.4) can, of course, also be given and define Poincaré covariant operations in $M\{I\}$. For example, $P^{\alpha\beta} + aI^{\alpha\beta}$ and $P^{\alpha\beta} - Q^{\alpha\beta}$ represent, respectively, a hypersphere center P and the hyperplane bisecting PQ orthogonally. (These twistors are not simple, so they represent "hyperspheres" rather than points.) We can check the coefficients in (10.4) using

$$P^{\alpha\beta}Q_{\alpha\beta} = -(p_i - q_i)(p^i - q^i), \tag{10.6}$$

which is a familiar Poincaré invariant expression.

The derivation of expressions such as (10.6) is facilitated if we specialize our twistor structure still further by the introduction of an *origin twistor* $O^{\alpha\beta}$ which represents a particular point O (the "origin") in M. Here, $O^{\alpha\beta}$ is to be simple, skew-symmetric, real and normalized as in (10.3)

$$O^{\alpha\beta}O_{\alpha\gamma} = 0, \qquad \bar{O}^{\alpha\beta} = O^{\alpha\beta}, \qquad O^{\alpha\beta}I_{\alpha\beta} = 2. \tag{10.7}$$

By putting $Q^{\alpha\beta} = O^{\alpha\beta}$ in (10.4) and (10.6), we obtain the twistor expressions for the Lorentz covariant vector operations of linear combination and squared magnitude.

Since the transformation group leaving both $I^{\alpha\beta}$ and $O^{\alpha\beta}$ invariant is the Lorentz group, we can expect to be able to express *spinors*, referred to the origin O, in *terms* of twistors. In effect, we carry out the construction given in Secs. II and III for twistors in terms of spinors, but in reverse. This gives us a correspondence between spinor indices and "reduced" twistor indices for which

$$\begin{aligned} O^{\alpha\beta} &\leftrightarrow \epsilon^{AB}, & I_{\alpha\beta} &\leftrightarrow \epsilon_{AB}, \\ O_{\alpha\beta} &\leftrightarrow \epsilon^{A'B'}, & I^{\alpha\beta} &\leftrightarrow \epsilon_{A'B'}. \end{aligned} \tag{10.8}$$

The *orthogonal idempotents* which reduce the twistor space—to a direct sum of two spinor spaces (one unprimed and one primed)—are

$$\begin{aligned} J^{\alpha}_{\beta} &= O^{\alpha\gamma}I_{\beta\gamma} \leftrightarrow \delta^{A}_{B} = \epsilon^{AG}\epsilon_{BG}, \\ \bar{J}^{\alpha}_{\beta} &= O_{\beta\gamma}I^{\alpha\gamma} \leftrightarrow \delta^{B'}_{A'} = \epsilon^{B'G'}\epsilon_{A'G'}. \end{aligned} \tag{10.9}$$

We have [from (10.1), (10.2), (10.7), etc.]

$$\begin{aligned} J^{\alpha}_{\beta}J^{\beta}_{\gamma} &= J^{\alpha}_{\gamma}, \qquad J^{\alpha}_{\beta}\bar{J}^{\beta}_{\gamma} = 0 = \bar{J}^{\alpha}_{\beta}J^{\beta}_{\gamma}, \\ \bar{J}^{\alpha}_{\beta}\bar{J}^{\beta}_{\gamma} &= \bar{J}^{\alpha}_{\gamma}, \qquad J^{\alpha}_{\beta} + \bar{J}^{\alpha}_{\beta} = \delta^{\alpha}_{\gamma}. \end{aligned} \tag{10.10}$$

A general twistor $P^{\alpha\beta\cdots\delta}_{\rho\cdots\tau}$, of valence $[{}^{r}_{s}]$ corresponds to a set of 2^{r+s} spinors. To obtain such a spinor, each index of $P^{\alpha\beta\cdots\delta}_{\rho\cdots\tau}$ is transvected with either a J^{α}_{β} or a $\bar{J}^{\alpha}_{\beta}$. The resulting twistor then represents a spinor with an unprimed index corresponding to each free J^{α}_{β} index and a primed index, in the *reverse position*, corresponding to each free $\bar{J}^{\alpha}_{\beta}$ index. For example,

$$P^{\kappa\lambda\mu}{}_{\varphi\chi}J^{\alpha}_{\kappa}\bar{J}^{\beta}_{\lambda}J^{\gamma}_{\mu}\bar{J}^{\varphi}_{\rho}J^{\chi}_{\sigma} \leftrightarrow \Pi^{A}{}_{B'}{}^{GR'}{}_{S}.$$

[38] In practice, the X^{α} and Y^{α} of the decomposition (9.3) (which may be specilized if desired) often turn out to be more convenient coordinates than $P^{\alpha\beta}$. This is of value in connection with physical fields and will be discussed elsewhere.

With the different possibilities for J^{α}_{β} and $\bar{J}^{\alpha}_{\beta}$, these give 2^{r+s} different spinors, which [by (10.10)] together determine $P^{\alpha\beta\cdots\delta}_{\rho\cdots\tau}$. Any spinor operation can thus be mirrored in twistor terms, using J^{α}_{β} and $\bar{J}^{\alpha}_{\beta}$. The correspondences of (10.8) and (10.9) are consistent with this.

The basic relations (2.13) and (3.1) are expressed as

$$X^{\beta}J^{\alpha}_{\beta} \leftrightarrow \xi^{A}, \qquad X^{\beta}\bar{J}^{\alpha}_{\beta} \leftrightarrow \eta_{A'},$$
$$L^{\beta}J^{\alpha}_{\beta} \leftrightarrow \lambda^{A}, \qquad L^{\beta}\bar{J}^{\alpha}_{\beta} \leftrightarrow \mu_{A'}.$$

If we put

$$\nu P^{\alpha\beta} = X^{\alpha}L^{\beta} - L^{\alpha}X^{\beta}$$

[cf., (9.3)] with ν chosen so that (10.3) holds, we get $\nu = \xi_{A}\lambda^{A}$. Then,

$$\left.\begin{aligned} P^{\rho\sigma}J^{\alpha}_{\rho}J^{\beta}_{\sigma} &\leftrightarrow \epsilon^{AB}, \\ P^{\rho\sigma}J^{\alpha}_{\rho}\bar{J}^{\beta}_{\sigma} &\leftrightarrow -ip^{A}_{B'}, \\ P^{\rho\sigma}\bar{J}^{\alpha}_{\rho}\bar{J}^{\beta}_{\sigma} &\leftrightarrow -\tfrac{1}{2}p_{i}p^{i}\epsilon_{A'B'}, \end{aligned}\right\} \qquad (10.11)$$

where $p_{AB'}$ is given as in (3.10). In the C picture P is the line joining two points X and L; in M^{*}, P is the (complex) point of intersection of two ("complexified") null lines. Thus, according to (3.10), p^{i} represents the *position vector*, with respect to O, of the point P in M^{*}. We can express (10.11) in matrix form as

$$(P^{\alpha\beta}) = \begin{bmatrix} \epsilon^{AB} & -ip^{A}{}_{B'} \\ ip_{A'}{}^{B} & -\tfrac{1}{2}p_{i}p^{i}\epsilon_{A'B'} \end{bmatrix}.$$

Expressions such as (10.6) and (10.4) then follow at once.

ACKNOWLEDGMENTS

The author is greatly indebted particularly to I. Robinson, but also to E. L. Schücking and to R. P. Kerr, for stimulating ideas in the early stages of this work.

This work was done partly at the University of Texas, where it was supported, in part, by the U. S. Air Force under Contract, AF-AFOSR-454-63.

30
CONSERVED QUANTITIES AND CONFORMAL STRUCTURE IN GENERAL RELATIVITY

Relativity theory and astrophysics I: Relativity and cosmology, lectures in applied mathematics **8** (1967), 147–159 (Ed. J. Ehlers, American Mathematical Society, Providence)

See Chapter 23.

756 Conserved quantities and conformal structure in general relativity. *Relativity theory and astrophysics I: Relativity and cosmology, lectures in applied mathematics* **8** (1967), 147–159

Conserved Quantities and Conformal Structure in General Relativity

1. **Conserved quantities in the presence of gravitational radiation.** In the linearized theory of general relativity, gravitational radiation is accompanied by the conservation of mass, momentum and angular-momentum for regions surrounded by empty space. However, as was first shown by Bondi [1] (cf., also [2], [9], [5]), mass loss will accompany outgoing gravitational radiation in the full theory. There are, nevertheless, for asymptotically flat space-times in the presence of gravitational radiation, ten *new* conserved quantities [4]. The sketch (Figure 1) shows how these and mass might be measured.

Here m_1 is the mass of the source before radiation and m_2 is the mass of the source after radiation. We may measure m_1 by examining the field at large distances from the source, on a space-like hypersurface which cuts the world tube of the body before it emits the radiation. If we try to measure m_2 similarly on an ordinary spacelike hypersurface which cuts this world tube after the radiation is emitted, then we find that the hypersurface must intercept the *radiation* also. Thus, instead of measuring m_2 we will be measuring

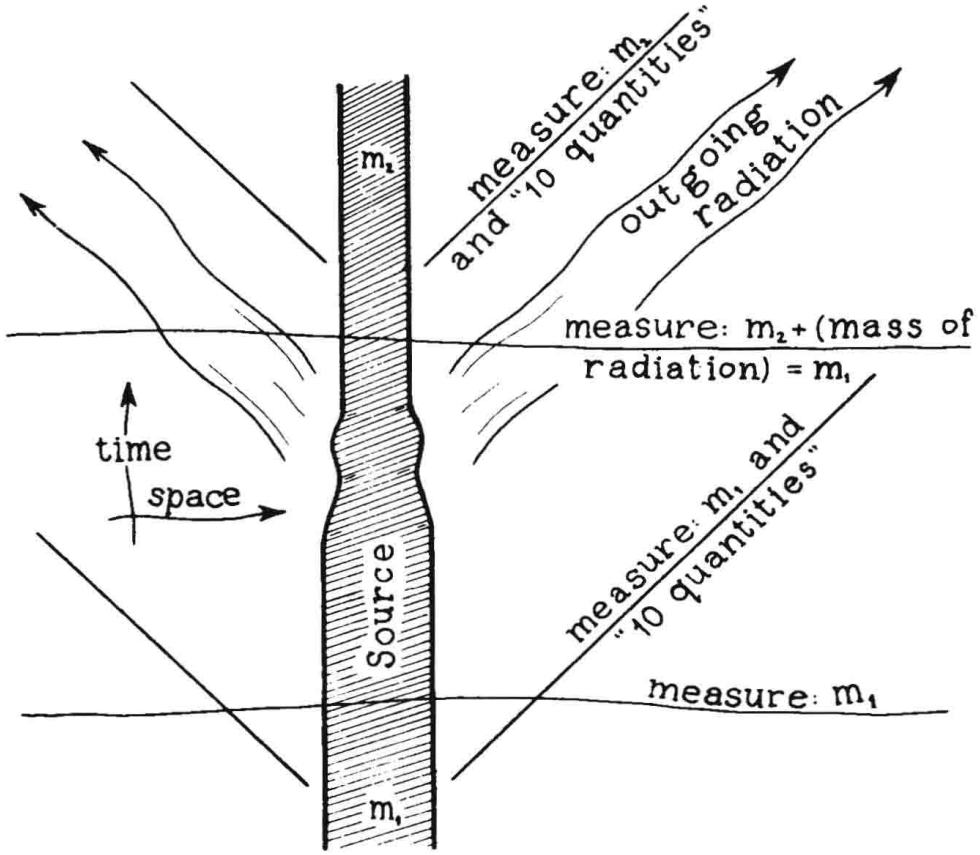


FIGURE 1

m_2 plus the mass carried by the radiation, which is again equal to m_1. We can avoid this difficulty if we use *null* hypersurfaces on which to measure mass. The mass m_2 can be measured if we choose a null hypersurface which lies entirely within the cone of radiation. We can similarly measure m_1 on a null hypersurface which opens out into the future before the radiation is emitted.

Similarly to the case of mass, the ten new quantities can be measured at infinity on any null hypersurface opening into the future, but unlike the case of mass they turn out to be the *same* whichever hypersurface is chosen provided the space-time is empty at infinity. Their exact definition will be given later.

The ten quantities can also be defined in the linear theory, but then it turns out that they always vanish for retarded fields.

For any *stationary* vacuum field (with asymptotic flatness) the ten quantities can be expressed schematically as

(1) (mass) × (quadrupole moment) − (dipole moment)^2.

(We will not prove this here, however.) The definition of these multipole moments can be taken from linear theory. In the full theory, both the source *and* its field must be expected to contribute to the values of the moments, but in any case, the quadrupole term measures an *asymmetry* of the source. The dipole moment includes an imaginary part describing the angular momentum. The expression is origin independent (as it must be since, if we use a null cone on which to measure the ten quantities, we can choose any origin as its vertex).

Let us do a simple thought experiment. We start with two hemispheres with their flat surfaces held apart by a small rod, as a stationary solution in asymptotically flat space-time. Then in a finite time interval this configuration is allowed to collapse into a sphere. As the gravitational field of the body changes, this is accompanied by a disturbance travelling outwards with the speed of light, so the body radiates gravitationally. It is easy to arrange things so that the source does not bounce (Figure 2).

Now, the ten quantities must be conserved. If our null hypersurface is chosen correctly, our measurement of the quadrupole and dipole moments of the sphere can surely be made zero so that the ten quantities would apparently vanish. But the ten quantities do *not* all vanish for our *original* configuration. Clearly this shows that the original configuration cannot be deformed into an *exact* stationary Schwarzschild solution under these conditions. Thus, it must be that more terms, which have to be added to the expression (1) to deal with *nonstationary* cases, are relevant here, so that the resulting ten quantities can be conserved. These terms perhaps involve an incoming field (since in the linear theory the ten quantities are nonzero only for fields with an incoming component). Here, this field would have to come from back-scattering of the gravitational waves, since we are supposing that no waves come in from infinity.

Note that in the Schwarzschild solution the curvature behaves as the mass times r^{-3}, so that we expect to measure mass in general by examining r^{-3} terms in the curvature. On the other hand, the ten quantities are obtained from r^{-6} terms in the curvature. The back-scattering presumably decreases locally with time even though the r^{-6} part of the back-scattered field does not decrease. One can think of the final state as being asymptotically spherical.

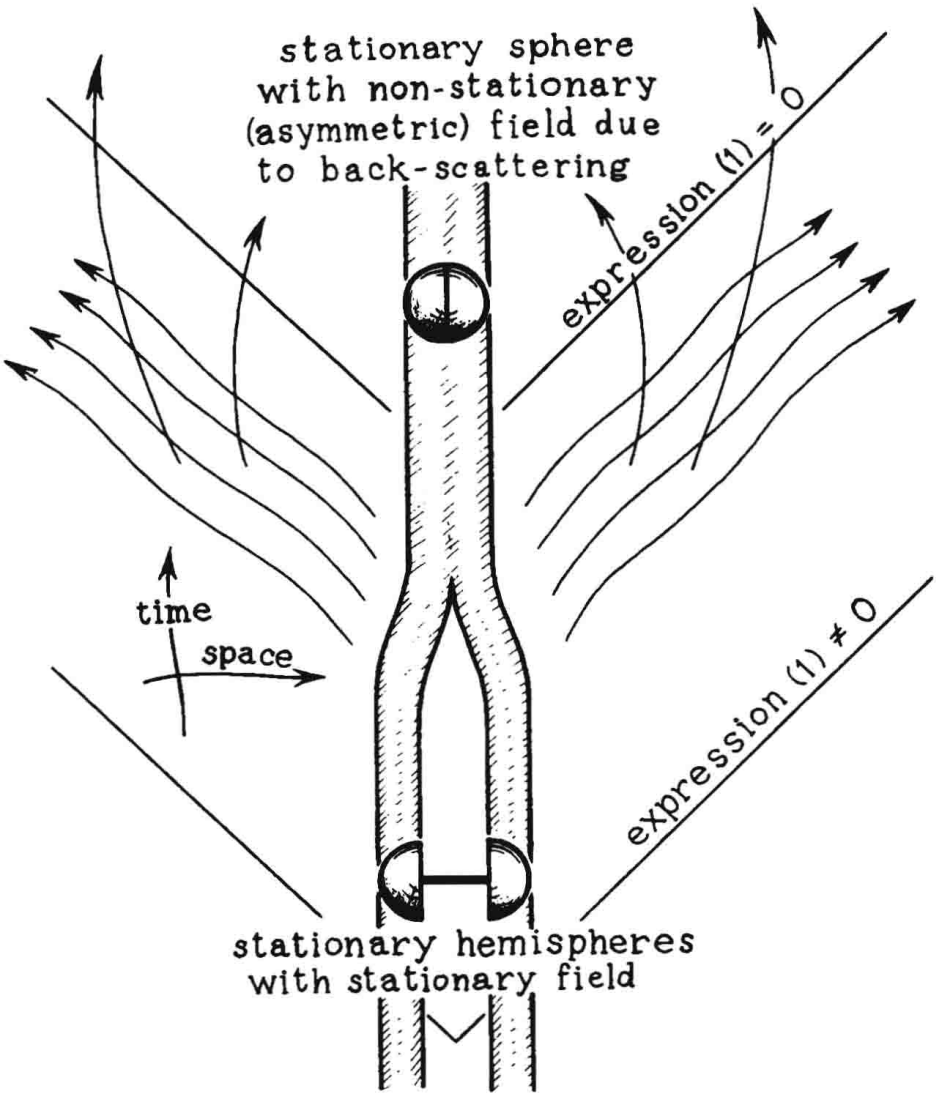


FIGURE 2. Back-scattering

2. **Conformal technique.** We wish to measure quantities at infinity and to set boundary conditions at infinity which insure the finiteness of various physical quantities (such as total energy, etc.). If we examine space-time from the point of view of its conformal structure only (e.g., angles, null cones, null geodesics), we can discuss infinity without recourse to an asymptotic limit. It turns out that we can, for asymptotically flat space-times, introduce two null hypersurfaces at infinity (past and future) which can be treated, con-

formally, the same way as finite hypersurfaces. Null infinity of the original space is represented by these hypersurfaces [6], [7].

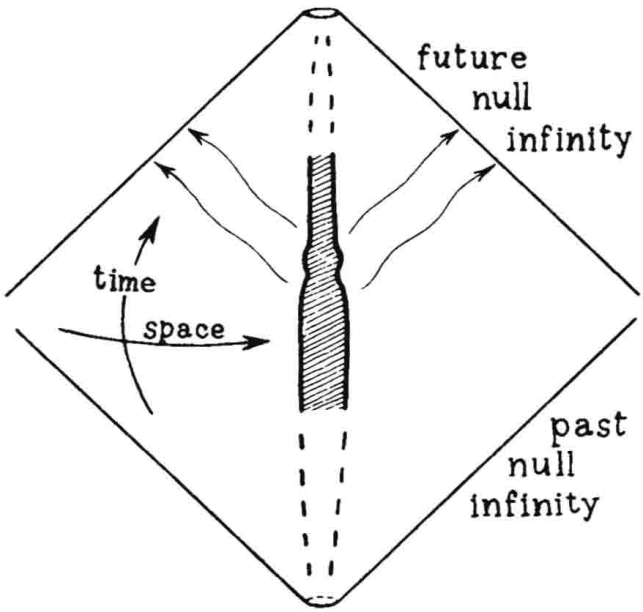


FIGURE 3

Let us consider the simplest case—Minkowski space-time $\tilde{M}_4$, where

$$d\tilde{s}^2 = du\,dv - (u - v)^2(d\theta^2 + \sin^2\theta d\phi^2)/4,$$

$$v = t + r, \qquad u = t - r.$$

We introduce a conformal transformation with conformal factor

$$\Omega = (1 + u^2)^{-1/2}(1 + v^2)^{-1/2}.$$

Here, tildes will always refer to quantities in the original physical space. The transformed space is a conformally equivalent, "non-physical" space M_4 with metric ds.

We put

$$ds = \Omega\, d\tilde{s},$$

and, since $\Omega \to 0$ at infinity, distance will become infinitely compressed at infinity. Now,

$$ds^2 = \Omega^2 d\tilde{s}^2 = dp\,dq - \sin^2(p - q)(d\theta^2 + \sin^2\theta d\phi^2)/4$$

with $u = \tan p$ and $v = \tan q$ (so that points at infinity are not given infinite coordinates). The two hypersurfaces representing infinity are given by $q = \pi/2$, $-\pi/2 < p < \pi/2$ and $p = -\pi/2$, $-\pi/2 < q < \pi/2$. These hypersurfaces represent the future and past null infinities. In addition there are three points I^- $(p = -\pi/2,\ q = -\pi/2)$, I^0 $(p = -\pi/2,\ q = \pi/2)$, and I^+ $(p = \pi/2,\ q = \pi/2)$ representing past timelike, spacelike and future timelike infinities. (I^0 is, in fact, a point, but this is not important here.) Space is regular in the neighborhood of the bounding hypersurface in M_4.

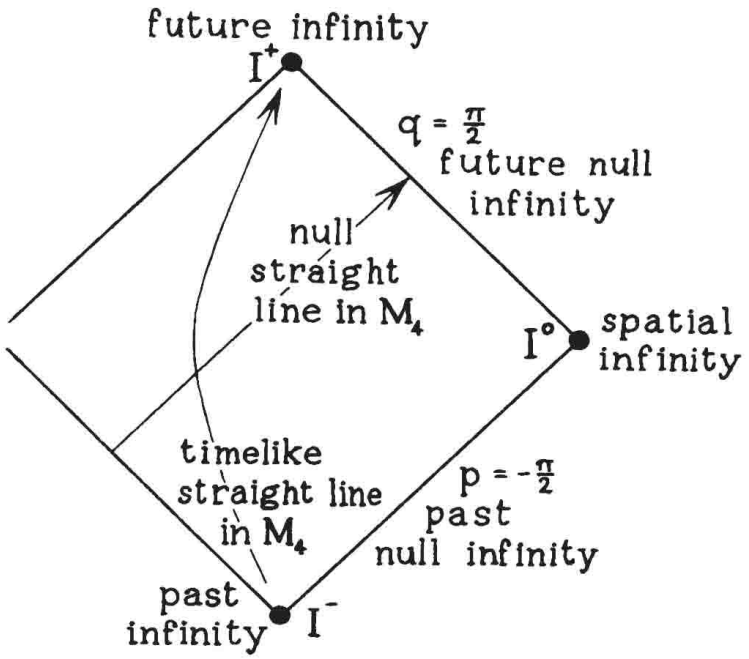


FIGURE 4. Finite M_4

Our definition of asymptotic flatness in curved space-time $\tilde{M}$ with $G_{\mu\nu} = 0$ at infinity is that M exist, with interior conformally equivalent to $\tilde{M}$ via $ds = \Omega d\tilde{s}$, so that:

(b) Ω suitably differentiable throughout M with $\Omega = 0$ on the boundary $\dot{M}$ of M,

(c) suitable topological conditions (e.g., null geodesics have two end points on $\dot{M}$).

The metric

$$d\tilde{s}^2 = R^{-2}A dR^2 + 2B_i dx^i dR + R^2 C_{ij} dx^i dx^j,$$

where $i, j = 1, 2, 3$ and A, B_i and C_{ij} are sufficiently differentiable

functions of x^μ ($\mu = 0, 1, 2, 3$), with $x^0 = R^{-1}$, includes as special cases many metrics which have been suggested as representing gravitational radiation [2], [8], [9] (and probably also [5]). If

$$\begin{vmatrix} A & B_j \\ B_i & C_{ij} \end{vmatrix} \neq 0,$$

this implies that

$$ds = \Omega d\tilde{s},$$

with $\Omega = R^{-1}$, is regular on our bounding hypersurface of M. (These metrics also include de Sitter space and the asymptotically de Sitter spaces of Demianski. These are not asymptotically flat in the usual sense, even though (a), (b), (c) are satisfied. This is because, here, $G_{\mu\nu} = -\lambda g_{\mu\nu} \neq 0$.)

3. **The field at infinity and the conserved quantities.** A zero rest-mass free-field can be regarded as being conformally invariant. For a spin two field consider a tensor $K_{\mu\nu\rho\sigma}$ subject to

$$\begin{aligned} K_{\alpha\beta\rho\sigma} &= K_{[\rho\sigma][\alpha\beta]}, \\ K_{\alpha[\nu\rho\sigma]} &= 0, \\ K^{\alpha}{}_{\rho\alpha\sigma} &= 0 \end{aligned} \tag{2}$$

and the "free-field equation"

$$\nabla_{[\mu} K_{\alpha\beta]\rho\sigma} = 0, \tag{3}$$

where ∇_μ indicates covariant differentiation (and the square brackets denote skew-symmetrization). In flat space, (2) and (3) is the gauge invariant description of a spin two, zero rest-mass free-field (i.e., linearized source-free gravitation). In curved space, the Equations (2) and (3) can still be used, but this implies a relation[1] connecting $K_{\mu\nu\rho\sigma}$ with the conformal curvature that must be satisfied. Equation (3) is preserved under the conformal transformation

$$g_{\mu\nu} = \Omega^2 \tilde{g}_{\mu\nu} \tag{4}$$

if

$$K_{\mu\nu\rho\sigma} = \Omega \tilde{K}_{\mu\nu\rho\sigma}. \tag{5}$$

[1] The general relation is $K_{\mu\nu\rho[\sigma} C_{\tau]}{}^{\mu\nu\rho} = 0$.

The conformal tensor $C_{\mu\nu\rho\sigma}$ satisfies (2), and also (3), whenever $G_{\mu\nu} = 0$. (Actually weaker conditions will suffice.) Thus, in the *original* space we can set

$$\widetilde{K}_{\mu\nu\rho\sigma} = \widetilde{C}_{\mu\nu\rho\sigma}, \tag{6}$$

since this will imply that (3) holds in $\widetilde{M}$, in the regions free of matter. However,

$$C_{\mu\nu\rho\sigma} = \Omega^2 \widetilde{C}_{\mu\nu\rho\sigma}$$

(cf., Schild, *Lectures on general relativity theory*, these Lectures). Thus, in the unphysical space this entails

$$C_{\mu\nu\rho\sigma} = \Omega K_{\mu\nu\rho\sigma}. \tag{7}$$

We now have the free-field Equations (3) holding for $K_{\mu\nu\rho\sigma}$ (and not $C_{\mu\nu\rho\sigma}$) in the unphysical space M, in the regions corresponding to the empty regions of $\widetilde{M}$. We wish to study the asymptotic gravitational field[2] by examining $K_{\mu\nu\rho\sigma}$ on the boundary hypersurface $\dot{M}$. For this to work, $K_{\mu\nu\rho\sigma}$ must be regular and finite on $\dot{M}$. But from (7), this clearly could only happen if $C_{\mu\nu\rho\sigma} = 0$ on $\dot{M}$. Fortunately, it can be shown that our assumptions already imply $C_{\mu\nu\rho\sigma} = 0$ there, and it follows that $K_{\mu\nu\rho\sigma}$ is *necessarily* regular on $\dot{M}$.

We will measure the energy-momentum and the ten conserved quantities on a null hypersurface $\widetilde{S}$, opening into the future, since we are considering outgoing radiation. Let S be the image of $\widetilde{S}$ in M. We are concerned with measurements made at infinity, so it is really only the quantities defined on the intersection Σ of $\dot{M}$ with S that will concern us. Suppose, now, we wish to compare a quantity measured on one hypersurface $\widetilde{S}_1$ with the same quantity measured on a "later" hypersurface $\widetilde{S}_2$. Let Σ_1 and Σ_2 be the respective intersections of M with S_1 and S_2 (Figure 5). We will have an integral over Σ_1 which we wish to compare with a corresponding integral over Σ_2. The idea here is to use Gauss' theorem on the three-surface $\dot{M}$. The difference between the integrals over Σ_1 and Σ_2 will then be expressed as an integral of some "divergence" over

[2] The point of view here is that $K_{\mu\nu\rho\sigma}$, subject to (2) and (3), is the zero rest-mass spin two field which describes "free gravitation," this "free gravitational field" being conformally invariant. The "interactions" are expressed through the Bianchi and Ricci identities as applied to $C_{\mu\nu\rho\sigma} = \Omega K_{\mu\nu\rho\sigma}$. The conformal *noninvariance* of these "interactions" is manifested in the explicit appearance of Ω in these expressions.

the portion M' of M lying between Σ_1 and Σ_2. If this divergence vanishes, we will have an exactly conserved quantity. Otherwise, we may be able to regard the integral over $\dot{M}'$ as the contribution to the quantity in question which is carried away by the gravitational radiation.

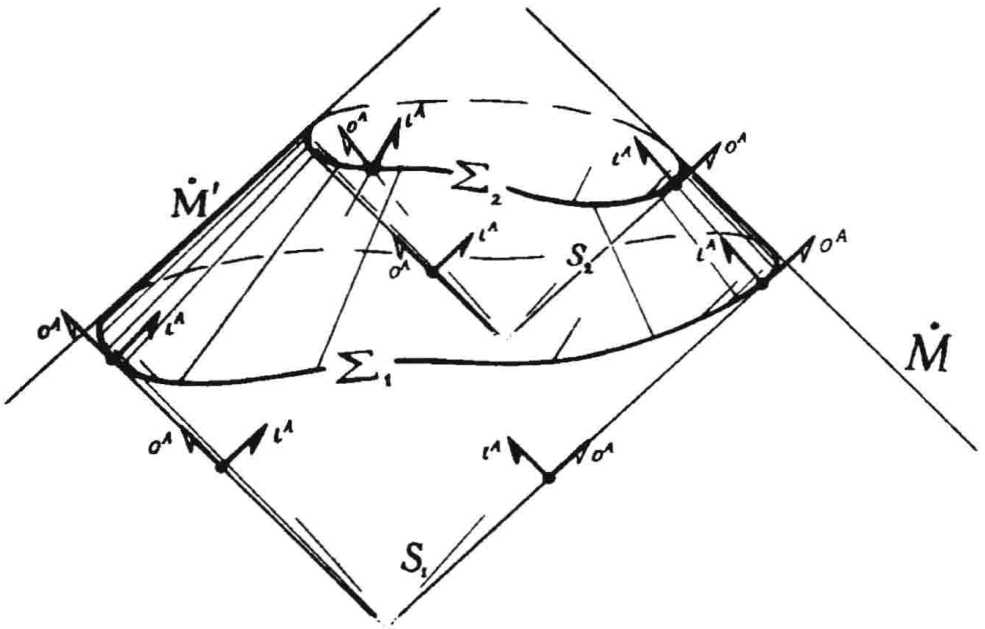


FIGURE 5

For asymptotically flat spaces, $\dot{M}$ is in fact a null hypersurface [7]. The structure of $\dot{M}$ is essentially the same as for Minkowski space (Figure 4). We shall omit the three points I^-, I^0, I^+ here. Then $\dot{M}$ consists of two portions, each of which is topologically a "cylinder" $S^2 \times \mathrm{E}^1$. We are concerned, here, only with the future portion $\dot{M}^+$, and by judicious choice of conformal factor Ω, we can ensure that the geometry of $\dot{M}^+$ is as simple as possible. In fact, by taking one generator of $\dot{M}^+$ "back to infinity" we can open out the cylinder into a space with Euclidean three-space topology. Furthermore, it turns out that we can also make this three-space metrically *flat* (Figure 6). This will simplify matters considerably.

In order to proceed further, we will need to use two-component spinors (cf., Schild, *Lectures on general relativity theory,* these Lectures). The quantity $\sigma_\mu{}^{AB'}$ (Hermitian (2×2) matrix for each μ) translates tensor quantities into spinor quantities and satisfies

$$\sigma_\mu{}^{AB'} \sigma_\nu{}^{CD'} \epsilon_{AC} \epsilon_{B'D'} = g_{\mu\nu}.$$

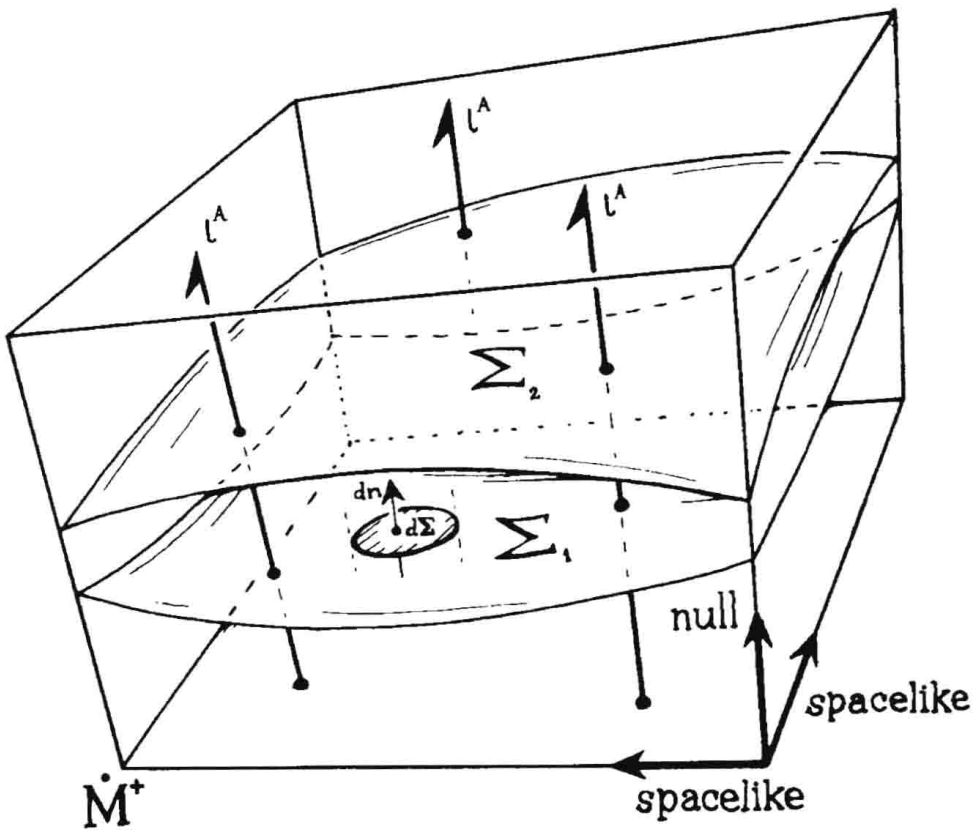


FIGURE 6

To preserve this under conformal transformation we set

$$\sigma_\mu{}^{AB'} = \Omega\tilde{\sigma}_\mu{}^{AB'} \tag{8}$$

with

$$\epsilon_{AB} = \tilde{\epsilon}_{AB}, \quad \epsilon_{A'B'} = \tilde{\epsilon}_{A'B'}.$$

We have

$$K_{\mu\nu\rho\sigma} \leftrightarrow K_{AP'BQ'CR'DS'} = \phi_{ABCD}\epsilon_{P'Q'}\epsilon_{R'S'} + \epsilon_{AB}\epsilon_{CD}\bar{\phi}_{P'Q'R'S'},$$

$$C_{\mu\nu\rho\sigma} \leftrightarrow C_{AP'BQ'CR'DS'} = \Psi_{ABCD}\epsilon_{P'Q'}\epsilon_{R'S'} + \epsilon_{AB}\epsilon_{CD}\bar{\Psi}_{P'Q'R'S'},$$

so that

$$\phi_{ABCD} = \Omega^{-1}\Psi_{ABCD} = \Omega^{-3}\tilde{\phi}_{ABCD} = \Omega^{-3}\tilde{\Psi}_{ABCD} \tag{9}$$

by (4), (5), (6), (7), (8).

We introduce a spinor reference dyad o^A, ι^A in the neighborhood of $\dot{M}^+$, with $o_A\iota^A = 1$ [3], where the null direction in $\dot{M}^+$ is represented by ι^A and the null directions in S_1 and S_2 are represented by o^A (Figure 5). It turns out we can choose ι^A and Ω so that

$$\left.\begin{aligned} &\nabla_{AB'}\iota_C = 0, \qquad o^A\bar{o}^{B'}\nabla_{AB'}o_C = 0 \\ &\Psi_{ABCD} = 0, \qquad R = 0 \\ &R_{AC'BD'}\iota^A\iota^B = -2N\bar{\iota}_{C'}\bar{\iota}_{B'}, \end{aligned}\right\} \text{on } \dot{M}^+ \tag{10}$$

where R is the scalar curvature and where $R_{AC'BD'}$ and $\nabla_{AB'}$ are, respectively, the spinor translation of the Ricci tensor $R_{\mu\nu}$ (which generally does not vanish in M) and of the covariant derivative operator ∇_μ. These equations will be required shortly.

Let $\tilde{\iota}^A$, $\tilde{o}^A$ and r be defined on $\tilde{M}$ so that the null directions in S_1 and S_2 are represented by $\tilde{o}^A$. Let both $\tilde{\iota}^A$ and $\tilde{o}^A$ be parallelly propagated along the generators of $\tilde{S}_1$ and $\tilde{S}_2$, and so that (near $\dot{M}^+$) ι^A, o^A are the respective images of $\tilde{\iota}^A$, $\tilde{o}^A$ in M. (This can always be arranged.) The radial parameter r is an affine parameter on each generator of $\tilde{S}_1$ and of $\tilde{S}_2$ (scaled so that $\tilde{o}^A\bar{\tilde{o}}^{B'}\nabla_{AB'}r = 1$). We have [7]

$$\tilde{o}^A \sim \Omega^{1/2}o^A, \qquad \tilde{\iota}^A \sim \Omega^{-1/2}\iota^A, \qquad r \sim \Omega^{-1}. \tag{11}$$

The five complex components of $\tilde{C}_{\mu\nu\rho\sigma}$ may be defined by

$$\begin{aligned} \Psi_0 &= \tilde{\Psi}_{ABCD}\tilde{o}^A\tilde{o}^B\tilde{o}^C\tilde{o}^D = O(r^{-5}), \\ \Psi_1 &= \tilde{\Psi}_{ABCD}\tilde{o}^A\tilde{o}^B\tilde{o}^C\tilde{\iota}^D = O(r^{-4}), \\ \Psi_2 &= \tilde{\Psi}_{ABCD}\tilde{o}^A\tilde{o}^B\tilde{\iota}^C\tilde{\iota}^D = O(r^{-3}), \\ \Psi_3 &= \tilde{\Psi}_{ABCD}\tilde{o}^A\tilde{\iota}^B\tilde{\iota}^C\tilde{\iota}^D = O(r^{-2}), \\ \Psi_4 &= \tilde{\Psi}_{ABCD}\tilde{\iota}^A\tilde{\iota}^B\tilde{\iota}^C\tilde{\iota}^D = O(r^{-1}). \end{aligned}$$

The order of magnitude behavior of these quantities is a consequence of the asymptotic flatness of the space ("peeling-off" property [3], [9]) and, more specifically, of the finiteness of $K_{\mu\nu\rho\sigma}$ (i.e., of ϕ_{ABCD}) on $\dot{M}^+$. This follows from (9) and (11) [7]. Put

$$\Psi_0 = \Psi_0^0 r^{-5} + \Psi_0^1 r^{-6} + O(r^{-7}), \tag{12}$$

$$\Psi_2 = \Psi_2^0 r^{-3} + O(r^{-4}). \tag{13}$$

By writing $\alpha_A = \iota^B\alpha_{AB}$, where α_{AB} is a symmetric spinor representing a bivector, we can put Gauss' theorem (applied to this bivector) into the form

$$\int_{\Sigma_1}\alpha_A o^A d\Sigma_2 - \int_{\Sigma_2}\alpha_A o^A d\Sigma_1 = \int_{\dot{M}'}\bar{\iota}^{C'}\nabla^A_{\ C'}\alpha_A\, d\Sigma\, dn,$$

where $d\Sigma_2$, $d\Sigma_1$, $d\Sigma$ are surface area elements ($d\Sigma$ referring to the

area of a cross-section of M'—the portion of M^+ between Σ_1 and Σ_2) and dn is an element of affine distance in the null direction in M^+ (Figure 6). For the ten conserved quantities we take

$$\bar{\alpha}_{P'} = o^A o^B o^C o^D o^E \nabla_{EP'} \phi_{ABCD} + 5 o^A o^B o^C o^D \phi_{ABCD} \nabla_{QP'} o^Q.$$

Then it turns out (from (10) and (3)) that

$$\bar{\iota}^{B'} \nabla_{AB'} \alpha^A = 0,$$

so that

$$\int_{\Sigma_1} \alpha_A o^A d\Sigma_1 = \int_{\Sigma_2} \alpha_A o^A d\Sigma_2.$$

Here

$$\bar{\alpha}_{P'} \bar{o}^{P'} = o^A o^B o^C o^D (o^E \bar{o}^{P'} \nabla_{EP'} + 5 \bar{o}^{P'} \nabla_{QP'} o^Q) \phi_{ABCD}. \tag{14}$$

In the physical space $\widetilde{M}$ this amounts to the term Ψ_0^1 of (12). (The $o^E \bar{o}^{P'} \nabla_{EP'}$ derivative acts along the generators of $\widetilde{S}_1$, $\widetilde{S}_2$ and takes us to the r^{-6} term. The term with $5\bar{o}^{P'} \nabla_{QP'} o^Q$ is a correction term to make the operation conformally invariant, and it knocks out the r^{-5} term of (12).) Thus, roughly speaking, if we integrate the Ψ_0^1 for $\widetilde{S}_1$, we get the same answer as for $\widetilde{S}_2$; however, we must take into account the weighting factors involved in the conformal transformation. In fact, we get different weighting factors depending upon which generator of $\dot{M}^+$ has been chosen to be "at infinity." Because of the conformal properties of (14), it turns out that there are just five linearly independent such weighting factors. The final result can be stated [4]:

$$Q_m = \int \Psi_0^1 {}_2\overline{Y}_{2,m} \sin\theta \, d\phi \, d\theta \qquad (m = -2, -1, 0, 1, 2),$$

where $r^2(d\theta^2 + \sin^2\theta \, d\phi^2)$ defines the metric of the sphere at infinity on $\widetilde{S}_1$ or $\widetilde{S}_2$. The "weighting factors" ${}_2Y_{2,m}$ are "tensor" spherical harmonics defined by

$${}_2Y_{2,m} = 24^{-1/2} \left(\sin\theta \frac{\partial}{\partial\theta} + i \frac{\partial}{\partial\phi} \right) \left(\frac{1}{\sin\theta} \frac{\partial}{\partial\theta} + \frac{i}{\sin^2\theta} \frac{\partial}{\partial\phi} \right) Y_{2,m}$$

where $Y_{2,m}$ are ordinary harmonics. The real and imaginary parts of the five Q_m are then the required conserved quantities.

In the same way, we can treat energy-momentum. Here, put

$$\alpha_A = -\phi_{ABCD}o^B\iota^C\iota^D + N\bar{\iota}^{C'}o^B\nabla_{AC'}o_B$$

(for N, see (10)). It then turns out that

$$\bar{\iota}^{C'}\nabla^A_{C'}\alpha_A = |N|^2.$$

The quantity N is essentially the Bondi-Sachs "news function," with $|N|^2$ representing the energy-momentum flux of the gravitational waves at infinity. We have

$$\alpha_A o^A = -\phi_{ABCD}o^A o^B\iota^C\iota^D + N\bar{\iota}^{C'}o^A o^B\nabla_{AC'}o_B$$

which, when translated into the physical space $\widetilde{M}$, becomes $-\Psi_2^0 + N\sigma^0$ (cf., (13)). The term $N\sigma^0$ is essentially a correction term arising from the fact that the hypersurfaces $\widetilde{S}_1$, $\widetilde{S}_2$ need not be as "isotropic" asymptotically as ordinary null cones in Minkowski space. The term σ^0 measures the asymptotic shear of $\widetilde{S}_1$, $\widetilde{S}_2$. Again the conformal transformation leads to different choices of weighting factors arising. Here there are four linearly independent choices—corresponding to "total energy" and the three components of "total linear momentum." The integral of $-\Psi_2^0$, "corrected" by $N\sigma^0$, over the sphere at infinity of $\widetilde{S}_1$ or $\widetilde{S}_2$ then gives the total mass of the system intercepted by the hypersurface.

References

1. H. Bondi, Nature **186** (1960), 535.
2. H. Bondi, M. G. J. van der Burg and A. W. K. Metzner, Proc. Roy. Soc. Ser. A **269** (1962), 21.
3. E. T. Newman and R. Penrose, J. Math. Phys. **3** (1962), 566; **4** (1963), 998.
4. ———, Phys. Rev. Lett. **15** (1965), 231.
5. E. T. Newman and T. W. J. Unti, J. Math. Phys. **3** (1962), 891.
6. R. Penrose, Lectures, Les Houches, 1963 Summer School of Theoret. Phys., Univ. Grenoble, pp. 563-584. Gordon and Breach, New York, 1964.
7. ———, Proc. Roy. Soc. Ser. A **284** (1965), 159.
8. I. Robinson and A. Trautman, Proc. Roy. Soc. Ser. A **265** (1962), 463.
9. R. Sachs, Proc. Roy. Soc. Ser. A **270** (1962), 103.

BIRKBECK COLLEGE, UNIVERSITY OF LONDON
LONDON, ENGLAND

31
THEORY OF QUANTIZED DIRECTIONS

Unpublished

This article, written in 1967, incomplete and unpublished, has nevertheless been quite widely cited, though circulated only privately in a handwritten form. The version appearing here has been transcribed, without change of content, specifically for this volume. I have added a brief Addendum at the end, outlining what I had been intending to include to complete the arguments. See also Chapters 28, 45, 46, 56, 69, and 110. For a published account of much of the material of this article, touching on its basic relation to loop-variable quantum gravity, see Seth Major, 1999, A spin network primer, *Am. J. Phys.* **67**, 972–80. A generalisation to the Poincaré group was made by Moussouris, J.P., 1983, *Quantum models of space-time based on recoupling theory*, Oxford D.Phil. Thesis, unpublished.

Theory of Quantized Directions

by Roger Penrose, August 1967
C/o Battelle Memorial Institute, 4000 N.E. Street, Seattle Wash. 98105

1. Introduction The concept of the continuum (or real number system) is one of the most fruitful mathematical notions employed in physical science. All successful and reasonably comprehensive physical theories that have so far been proposed have rested heavily on this notion, if only because space and time according to our present ideas, form a continuum and are therefore to be represented by continuous coordinates. And quantum mechanics (as normally understood), although it has introduced a measure of discretion into physical theory, rests even more heavily on the continuum concept than did classical physics. For the array of complex numbers is called upon at the outset in connection with probability amplitudes and the superposition law. Nevertheless, there would appear to be strong reasons for believing that the continuum concept may eventually have to be abandoned as one of the basic ingredients of a fundamental physical theory. A theory could, perhaps, be based on purely combinatorial principles and only at a second stage, in the role of an approximation, or of a computational or conceptual aid, would the continuum emerge. In this way, it has sometimes been suggested, the divergences and inconsistencies which plague quantum field theory might possibly be avoided.

Perhaps a significant reason for doubting that the continuum occupies a basic position in fundamental physics is a belief in the ultimate simplicity of nature. From a logical point of view, the continuum is not at all a simple structure. Also, conceptual difficulties arise in the definition of a real number which are of quite a different order from the case of an integer (or a fraction). It is interesting to recall, in this context, Kronecker's famous dictum: "God made the integers, all the rest is the work of man"! (Ref. 1.) These difficulties have sometimes prompted various writers (Ref. 2) to try to replace the usual notion of the continuum by some other structure of a more constructive nature which would avoid these conceptual problems. However, it appears that a price paid is always an increase in the complexity of the structure involved.

As has been pointed out by Schrödinger, Science & Humanism (Ref. 3) the continuum only seems to be a relatively simple idea because it is so familiar. The idea has developed historically from a mathematical idealization of a physical "straight line". Implicit in this is the assumption that such a "line" would appear similar no matter how many times it is magnified. But as we now know, the most precisely drawn material line would, on account of its atomic structure, necessarily appear totally different if magnified by a factor of, say, 10^{10}. It might be maintained, however, that the atomistic nature of matter in no way argues against continuity for a background space. But the fact remains that at the level at which continuity fails for matter, our initial reason for believing in its applicability to space itself, now becomes of doubtful value. The present reasons for believing in the continuity of space and time amount to the fact that theories based on such a continuity appear to work. On the other hand this should not necessarily be taken as an indication that space and time "are" continuous at the level under consideration. It should be recalled that Newton's Theory of gravitation which operates in a flat space and

uses the concept of a gravitational force also "works". But these Newtonian ideas must be thought of as "wrong", even at the level of accurate applicability, when viewed from the conceptually more satisfying framework of general relativity. Furthermore Schrödinger argues forcefully that discontinuities are present in physical phenomena at the quantum level whereas it is quantum theory (i.e. wave mechanics) that assumes continuity is maintained right up until the final step at which the "observation" is made.

What will be proposed here is a model theory whose rules are essentially simple and entirely combinatorial in character — and from which the beginnings of space-time geometry emerge when structures of sufficient complexity are considered. At this stage no attempt is made at modifying the content of physical theory. Only a reformulation is suggested in which quantities normally requiring continuous coordinates for their description are eliminated from primary consideration. In particular, since space and time have therefore to be eliminated, what might be called a form of Mach's principle must be invoked: only the relationships of objects to each other can have significance — a relationship of an object to some background space should not be considered.

What physical quantities, then, can be used as a starting point? In the first instance such a quantity must have a discrete spectrum preferably taking values which are integer multiples of some elementary value. Secondly, we must expect that the rules for combining the quantities are essentially bound up with quantum processes. Finally, since we expect to be able to build up the idea of space from these rules, such a quantity must be intimately related to space-time properties — although it must be a scalar since space-time directions cannot appear at the outset. The one quantity which seems to emerge from all this is total angular momentum. A more satisfactory alternative might be to try to construct a theory using both rest-mass and total angular momentum as primary concepts. Unfortunately, however, too little is known about the mass spectrum at present, to indicate how to proceed, and there is no suggestion as yet that mass is based on integral multiples of an elementary mass.

In order to make progress, we shall suppose here that the universe is analyzable into substructures each of which may be momentarily considered as isolated from the rest of the universe and having a well-defined total angular momentum. These substructures will be called units. They could be elementary particles or composite particles such as nuclei, atoms, or perhaps molecules, etc. Associated with each unit is a non-negative integer called its spin number. The spin number is twice the total angular momentum quantum number of the unit (so that for an electron it is one, for a deuteron it is two, etc. A unit of spin number n will be called an n-unit. The only pieces of information about the universe that are allowed to be considered here are the network of interrelationships between the various units (i.e. the knowledge of which units combine or split up into which other units) and the associated spin numbers. Any such network with assigned spin numbers will be called a spin network. Laws of physics then reduce to statements about frequencies of occurrence in the universe of the various spin networks.

The idea here is to construct a geometry of directions in space from the laws of spin networks. Now, according to quantum theory, a system of zero total angular momentum must be spherically symmetrical. Dirac (Ref. 4). To a 0-unit, therefore, any direction in space is equivalent to every other. In effect, only one distinguishable "direction" in space appears to a 0-unit to exist. Furthermore, a 1-unit (e.g. an electron) can have only two orientations. Thus, only two external spatial "directions" exist for such

a unit. More generally, only n+1 possible "directions" appear to exist for any n-unit. Thus, a continuum of possible external directions can only be approached for units with large spin numbers. The suggestion is, therefore, to associate a direction — which may be thought of as the axis of spin—with each unit of large spin number. Such a unit will be called a large unit. It is possible to define angles between two such directions in the following way. An "experiment" is performed between two large units whereby a 1-unit is detatched from one of the large units and is then united with the other one. The relative probabilities that the spin numbers of the large units increase or decrease as compared with each other give a measure of an "angle" between their spin directions (consistent with the predictions of quantum mechanics). However, part of the probability may be due to "ignorance". This is manifested in insufficient connecting links between the two large units in the known spin networks. This possibility can be eliminated from consideration by means of a repetition of the above experiment. If the probabilities in the second experiment remain essentially unchanged by the result of the first experiment, then the "ignorance" factor is deemed to be insignificant and the "angle" between the large units well defined.

The main result of this paper is that if the laws for spin networks are taken to agree with the rules for combining non-relativistic angular momenta according to quantum mechanics, then the geometry of "directions" defined by large units as described above agrees, in the limit, with the geometry of directions in a three-dimensional Euclidean space. In order to obtain four-dimensional space-time, corresponding relativistic laws would have to be employed which take into account the possibility of a relative "velocity" between units — or else some other method would have to be employed. Also, with relative velocities present the mixing of orbital angular momentum with the spin should give rise to a notion of "distance" as well as of four-dimensional angle. With the present scheme distances do not emerge, and time is absent even to the extent that the time-ordering of events is totally irrelevant. But within the framework of the situations treated, the result obtained here is reassuring. In fact, in at least one respect it is somewhat remarkably so. For the initial assumption of our being able to associate a direction of spin with any quantum mechanical system of large total angular momentum is by no means obviously justified. Moreover, it would appear at first sight to be false! Wave functions of high total angular momentum need single out no particular direction as an axis of spin (e.g. a state with vanishing magnetic quantum number; furthermore, a state with spin number twenty, say, could have icosahedral symmetry, etc.). The reason the approach seems to work appears to be connected with the "Machian" nature of the theory: the geometry of directions that is built up is the one defined by the system itself and it may not agree with the geometry of some previously given (and unnecessary!) background space.

A second purpose of this paper is to introduce a new combinatorial calculus for treating quantum mechanical angular momenta. The results obtained are equivalent to a use of the conventional Clebsch-Gordan-Wigner symbols etc. and the method seems well suited to a discussion of problems such as the one treated here. The calculus is somewhat related to the graphical techniques of Ord-Smith, Edmonds, and Levinson (Ref. 5), but appears to have a greater flexibility. The basic rules are essentially very simple. It is also shown, here, how this calculus can be applied to obtain various known formulae. In particular, direct transformations are found which exhibit Regge's 3-j and 6-j symmetries.

Also, a formula is given which contains as special cases, Racah's explicit expressions for both the 3-j and 6-j symbols. (Ref. 6)

2. Spin Networks, Values and Norms. Any spin network can be represented as a diagram in which the units are depicted as line segments numbered according to their respective spin numbers (Fig. 1). At one or both ends of each line segment may be a vertex which represents an "interaction"; i.e. the coming together of two units to form a third or the splitting up of a unit into two other units. "Time ordering" is actually irrelevant but it may be thought of as proceeding from the bottom of the diagram to the top. In the spin networks considered here, exactly three line segments must terminate at each vertex and the total number of these units must be finite.

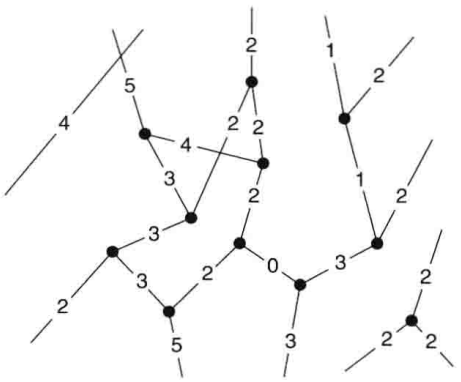


Fig. 1 An example of a spin network. The lines may be thought of as representing the world-lines of particles. The associated integer is twice the total angular momentum quantum number. The vertices represent interactions. Three edges must meet at each vertex.

Any unit which is not terminated at both ends by a vertex will be called an end unit. A spin network possessing no end unit will be called closed. An example of a closed spin network is given in Fig. 2. Any closed spin network determines, up to sign, a certain integer called its value, for which a combinatorial definition will be given shortly. The value can be positive or negative (or zero) and its sign is fixed once an orientation at each vertex is specified which determines a cyclic order for the three units which meet there. Once the spin network has been drawn on a plane such as in Fig. 2, this conveniently determines the orientations to be (say) anticlockwise at each vertex.

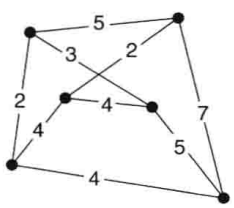


Fig. 2 A closed spin network. Every line segment connects two vertices.

For any open (or closed) spin network a non-negative integer is defined called its norm. The norm of a spin network is defined as the modulus of the value of the closed spin network which consists of two copies of the original network joined along their corresponding end units. An example illustrating this is given in Fig. 3. (When drawing

the diagrams on a plane, it is often convenient to represent the two networks as mirror images of each other. The implied orientations, of course, do not affect the norm.) The norm of a spin network α is denoted by $\|\alpha\|$, while for a closed oriented spin network drawn on a plane, the network itself denotes its value.

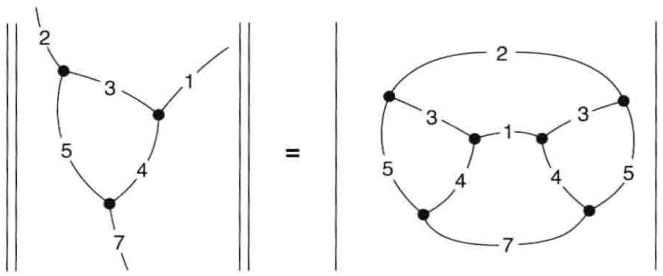


Fig. 3 To determine the norm of the spin network on the left, two copies are drawn side by side and connected along their free ends. Here, the networks are drawn as mirror images of each other. The modulus of the value of the resulting closed network is the required norm.

The importance of the concept of the norm of a spin network, here, lies in the fact that it can be used to determine the probabilities of the various spin numbers for a given network which has a suitable subset of its spin numbers already assigned. The rule will be stated in full generality here, but the proof is left until section 4. Effectively a special case of this rule will be the following.

(2.1) The norm of a spin network vanishes if and only if the physical situation it represents is forbidden by the rules of non-relativistic quantum mechanics.

For more general situations the norm gives, under appropriate circumstances, a measure of the frequency of occurrence in the universe, of various sets of unknown spin number values for a given spin-network configuration. Consider a network ω, a portion κ of which consists of units whose spin numbers are known. A vertex of ω is to belong to the (possibly disjointed) spin network κ whenever all three units associated with the vertex belong to κ. Thus, we can imagine the network ω to be "cut" across some of its edges into two parts, one of which is κ. Let ξ denote the remaining portion of ω. Thus, ξ is a spin network containing all the vertices of ω which are not in κ and units corresponding to those of ω which belong to these vertices. In order for the formula which follows to be applicable it will be necessary to assume that the network of ξ is a tree (i.e. a connected network containing no closed circuits of units) and that exactly one of the end units of ξ — say an x-unit — is an end unit of ω (and not of κ). All the other end units of ξ will then also be end units of κ. The situation is illustrated by an example in Fig. 4. The probability for the spin values assigned to ξ, given κ, is (assuming $\|\kappa\| \neq 0 \neq \|\xi\|$, cf. (2.1))

$$\text{probability} = \frac{(x+1)\|\omega\|}{\|\kappa\|.\|\xi\|} \tag{2.2}$$

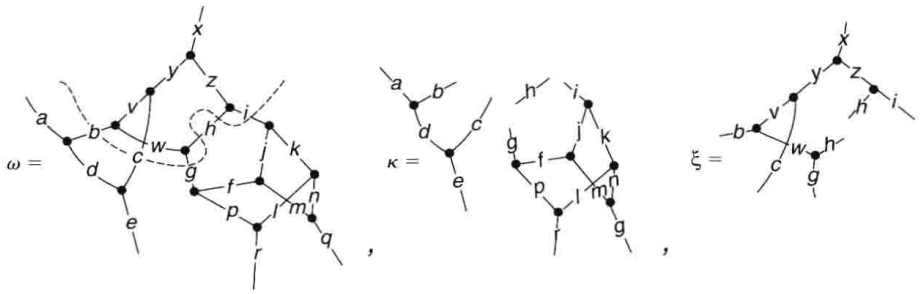


Fig. 4 The type of situation to which formula (2.2) applies. The network structure of ω is supposed known, as are the actual spin numbers for the (disjointed) portion κ. The portion ξ must be a tree, exactly one of whose end units is an end unit of ω. The probability for the set of values v, w, x, y is then determined by (2.2).

The special case when x is the only unknown in ξ will be the one of most interest to us here (see also (2.12)).

To define, combinatorially, the value of a closed oriented spin network we proceed as follows. (Proofs are left until section 4.) We imagine the line segment representing an s-unit to be composed of s parallel (non-oriented) strands. At each vertex the ends of the strands can be joined together in pairs in various ways. Each such way of connecting the strands results in a number of closed loops (see Fig. 5). Each closed loop is assigned the value –2. The value of r closed loops is then $(-2)^r$. Each such system of closed loops contributes linearly to the total value N of the network:

$$N = \Pi \frac{1}{s!} \times \sum \varepsilon (-2)^r \tag{2.3}$$

where ε is ±1 or zero, depending on how the strands are connected at the vertices and where, for later convenience, a factor 1/s! multiples the whole expression, for each s-unit in the network. In fact N remains an integer, although this is not immediately obvious at this stage and will emerge from the discussion of section 5. The definition of ε is determined up to a sign for the whole network by the following two rules.

(2.4) If, at any vertex, two strands belonging to the same unit are connected to each other, then $\varepsilon = 0$

(2.5) Otherwise, $\varepsilon = \pm 1$ where the sign of ε changes whenever the connections to two strand ends belonging to a unit are interchanged.

(See Fig. 5) This is in fact sufficient for the probability law (2.2), since the norm does not depend on the sign of the value. However, for the discussion to follow, a precise determination of the sign of ε is necessary. This is a little more complicated since it depends on the orientation assigned to each vertex.

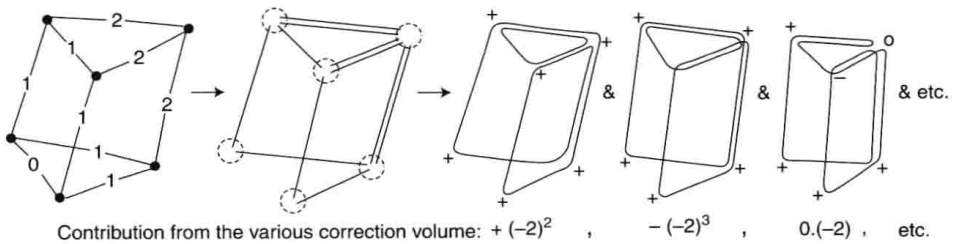


Fig. 5 Each line of the closed network on the left is represented as n strands, where n is the corresponding spin number. The ends of the strands are connected in pairs at each vertex in all possible ways. If r complete circuits emerge, the term $(-2)^r$ is to be associated with the connection scheme. As explained in Section 3, an additional sign is included for each crossing point, on the plane, of one strand connection with another which occurs at a vertex. If a strand doubles back on itself at a vertex, the term is zero.

Label the various units in the network A, B, C, D,..., where A is an a-unit, B is a b-unit, etc. The a strands belonging to A can be labelled consecutively from A_1 up to A_a, but for consistency with later conventions, the strand ends instead will be so labelled, where the labelling of the strand ends at one end of the unit runs in the opposite direction from those at the other end of the unit, that is, if the label is A_j at one end of the strand, it is A_{a-j+1} at the other end (j=1,..., a). The labeling of strand ends for B, C, ... follows similar conventions. Consider a vertex at which the units A, B and C come together and suppose the assigned orientation at the vertex specifies the cyclic order ABC for the units. Consider any connection scheme of a strand ends satisfying (2.4) given by pairings:

$$(A_gB_h) \ldots . (A_kB_l) (A_mC_n) \ldots .(A_qC_r) (B_sC_t) \ldots (B_wC_x). \tag{2.6}$$

Then we have $\varepsilon = +1$ or -1 according as the ordering in (2.6) is an even or odd permutation of

$$A_1A_2 \ldots . A_aB_l \ldots .B_bC_1 \ldots C_c. \tag{2.7}$$

It should noted that the ordering of the pairs in (2.6) is irrelevant, since the interchange of two pairs is always an even permutation. Furthermore, (2.5) is clearly satisfied by the fact that ε, so defined, depends only on the cyclic order of A, B, C which will emerge in a moment.

We observe that the number of strands connecting A to B plus the number of strands connecting A to C at the vertex must equal a. From this and the two similar relationships for B and C, we get:

$$\begin{aligned} &\frac{1}{2}(a+b-c) \text{ strands connect A to B at the vertex} \\ &\frac{1}{2}(b+c-a) \text{ strands connect B to C at the vertex} \\ &\frac{1}{2}(c+a-b) \text{ strands connect C to A at the vertex} \end{aligned} \tag{2.8}$$

These three expressions must represent non-negative integers, so for the possibility of a non-zero ε we must have the following two laws.

(2.9) The sum of the spin numbers of the three units coming together at any vertex must be even.

(2.10) The sum of the spin numbers of two units at a vertex cannot be smaller than the spin number of the third.

Once (2.9) and (2.10) are satisfied a connection scheme with $\varepsilon \neq 0$ will be possible.

The law (2.9) will be recognized as a statement of the familiar fact that the total number of fermions (in and out) involved in an interaction must be even. The law (2.10) is a well-known "triangle inequality" in the theory of angular momentum (no relative velocities!). The norm of any spin network not satisfying (2.9), (2.10) must vanish. This is to be expected in view of the physical interpretation (2.1). The spin networks of Figs. 1, 2 and 3 have all been drawn in accordance with (2.9) and (2.10).

To see that ε depends only on the cyclic order of A, B, C at the vertex, change the ordering ABC into BCA in (2.6) and (2.7). This results in a(b+c) interchanges in (2.7) and in a^2 interchanges in (2.6) since the ordering of the pairs is irrelevant. But a(b+c) and a^2 have the same parity because of (2.9). Hence ε remains unchanged, whence the sign of the value of any closed spin network is fixed once a cyclic orientation is assigned at each vertex. To see the effect of reversing the orientation at any one vertex, replace the ordering ABC in (2.6), (2.7) by BAC. We get ab interchanges in (2.7) and $\frac{1}{2}(a+b-c)$ interchanges in (2.6) (see (2.8)). Thus, the value of the closed spin network is changed by a factor

$$(-1)^{(a+b-c)/2-ab} = (-1)^{a(a-1)/2+b(b-1)/2+c(c-1)/2} \tag{2.11}$$

the second formula being obtained from the first by a little manipulation using (2.9). It shows, incidentally, that a reversal of the orientation of all the vertices in a closed spin network leaves its value unaltered.

In order to complete the definitions for values and norms of spin networks, one final convention is necessary.

(2.12) The value of the closed "spin network" which consists of a single s-unit forming a closed loop is $(-1)^s(s+1)$. Although this is, perhaps, not strictly a spin network in the sense intended initially, the definition is consistent with the other conventions, as will emerge later. Also, the definition is necessary for determining the norm of the (open) spin network which consists of a single s-unit, the norm being s+1. Any spin network (see κ in Fig. 4) which has an (open) s-unit as a disconnected part has, therefore a norm which is s+1 times the norm of the remainder.

One further fact that follows from (2.7) is that for $\varepsilon \neq 0$, the connection scheme at each vertex is unique up to a permutation of the strand ends belonging to each unit, since the number of strands from each unit to each other is fixed. It follows from this that rule (2.5) does, in fact, determine the sign of ε up to an overall ambiguity for the whole (unoriented) spin-network. In the next section a diagrammatic method for determining the sign of each term in (2.3) will be given, which depends on a representation of the spin network on a plane (Fig. 5). This somewhat simplifies the appearance of the definition of the value of a spin network and will be especially useful in relation to later calculations.

3. Diagramatic Notation: Binors Let α be an oriented spin network represented on a plane, as in Fig. 1, so that the orientation at each vertex is anticlockwise. The units are represented as smooth curves which may cross each other, in the diagram, at some points other than the vertices. These crossing points will be called spurious intersections as they have no physical significance. They are necessary in order for some networks (such as the one of Fig. 2) to be representable on a plane. The diagram should be drawn so that the spurious intersections do not occur at vertices. To represent each term in (2.3) we draw s parallel strands (as in Fig. 3) corresponding to each s-unit. No intersections except for the spurious ones are to occur between strands, until the connections are made at the vertices.

Each connection scheme at each vertex must be drawn (Fig. 6) according to the following two conventions.

(3.1) No connecting strand shall be self-intersecting

(3.2) All intersection points must be simple.

A simple intersection occurs when exactly two strands cross, with distinct tangents there.

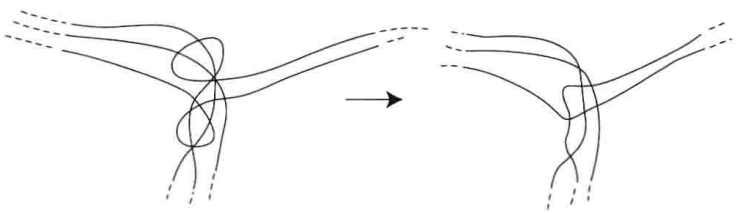


Fig. 6. The left-hand diagram contains self-intersecting connecting strands and non-simple intersections. These must be removed by altering the paths of the strands as on the right.

Having thus drawn a diagram for a given connection scheme which satisfies (2.4) (i.e. $\varepsilon \neq 0$) the rule for (2.3) is as follows:

(3.3) $\varepsilon = +1$ or -1 according as the total number of intersections between the strands—not counting spurious intersections—is even or odd.

To see that (3.3) agrees with that given by (2.6), (2.7), observe, first, that the sign given by (3.3) is independent of the particular way in which the connections are drawn provided the conventions (3.1) and (3.2) are adhered to in the drawing: simple topological considerations* show that allowed intersections can only appear or disappear in pairs. Now consider units A, B, C meeting at a vertex and draw the vertex with the unit lines entering from the bottom of the diagram in the order ABC. (Fig. 7) The strands can then be depicted entering from the bottom in the order $A_1A_2\ldots\ A_aB_1\ldots B_bC_1\ldots C_c$ (see (2.7)). This will be consistent with the labeling for the other vertices because of the rule that the labeling is in the reverse order at the other end of each unit. (This is necessary if the strands are not to cross over along the unit.) At the top of the diagram the pairings (2.6) can be represented as a row of hoops, the ends of which can be directly connected to the relevant strand ends $A_1,\ldots,C_c$. The part of the diagram which gives the connections between the hoop ends and the strand ends (Fig. 7) is an Aitken diagram for the required permutation — transforming (2.7) into (2.6). Each intersection point represents a simple

* There is a curious irony here that arguments based on properties of the mathematical continuum are being used to illucidate purely combinatorial scheme.

transposition, so it follows that the permutation is even or odd according as the number of intersections is even or odd (Ref. 7). Combining the results for all the vertices we arrive at the rule (3.3). (Ref. 8).

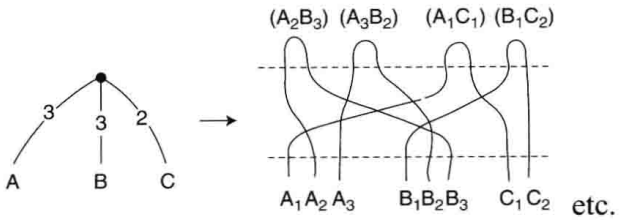


Fig. 7 A connection scheme for the vertex on the left is given on the right. The portion between the broken lines is an Aitken diagram for the permutation $A_1A_2A_3B_1B_2B_3C_1C_2C_3 \rightarrow A_2B_3A_3B_2A_1C_1B_1C_2$

Although (3.3) simplifies the appearance of the expression (2.3) for the value of a closed oriented spin network, this formula is still not very practical for evaluating anything but the simplest of closed spin networks. A very simple example is evaluated in Fig. 8. Some more practical methods will be developed in section 5 after the proofs are given in Section 4.

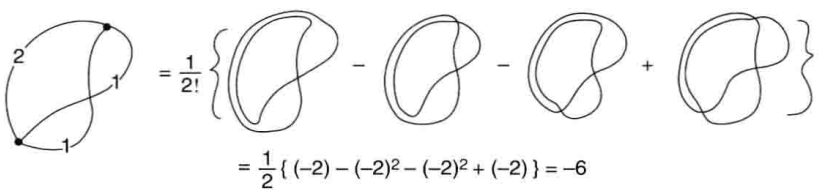


Fig. 8. The evaluation of a simple closed spin network according to the formula (2.3) and the rules (2.4) and (3.3). Any closed loop denotes the integer –2.

As a preliminary to this and for the various proofs, it will be convenient to develop the notation further. We may, in fact, regard any vertex, at which (say) an a-unit, a b-unit and a c-unit come together, as a formal sum, with the appropriate sign attached, of all the different possible strand connections for which $\varepsilon \neq 0$. The case a = 3, b = 2, c = 1 is illustrated in Fig. 9, where the sum consists of six terms. For the general case there will be

$$\frac{a!b!c!}{\left\{\frac{1}{2}(a+b-c)\right\}!\left\{\frac{1}{2}(b+c-a)\right\}!\left\{\frac{1}{2}(c+a-b)\right\}!} \tag{3.4}$$

terms in all (see 2.8). A formal sum of strand connection such as this, with integer (or perhaps rational) coefficients will be called a binor. A more complete definition of a binor will be given shortly.

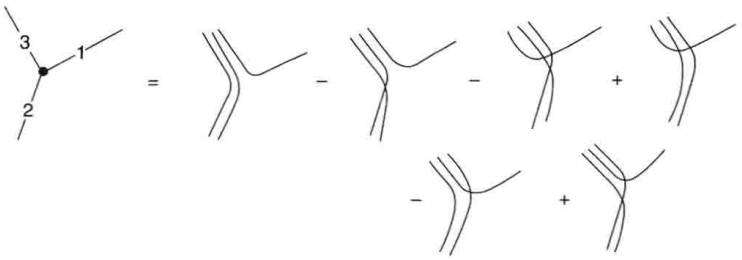


Fig. 9. The representation of a vertex as a formal linear combination of strand connection schemes. This formal sum is a simple example of a binor.

If all the vertices of a closed spin network are expanded (distributively) in this way and if each of the resulting closed strand loops is counted as –2, then according to formula (2.3), the result will be just Πs! times the value of the spin network (compare Fig. 8). In order to eliminate this factor, moreover, the notational convention will be employed that, in any spin network diagram, each s-unit which has a vertex at both ends denotes not just s strands, but s strands multiplied by the factor 1/s!. (For example, see Fig. 10.) It turns out that the resulting binor will still have integral coefficients. In the case of spin networks which are simple trees this is, in fact, fairly obvious, since each factor 1/s! removes a redundancy among the terms in the expansion (see Fig. 10). The general argument is left to section 5, however.

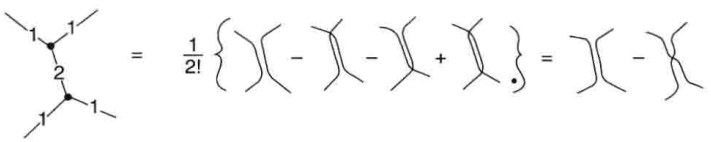


Fig. 10 The binor representing a spin network has a factor 1/s! incorporated for each s-unit which has a vertex at both ends. In the case of tree diagrams this removes a redundancy in the expansion.

For a more complete discussion of binors and their properties it will useful to introduce another notation which brings out a remarkable similarity between binor algebra and certain tensor algebras. A single strand starting from a point α and ending at a point β will be denoted by $S_{\alpha\beta}$. If α,β,γ,δ,…,μ are distinct points and strands are depicted connecting α to β, γ to δ, …, μ to ν, then the corresponding binor is $S_{\alpha\beta}\, S_{\gamma\delta}\ldots S_{\mu\nu}$. This is to be thought of as a product—the outer product—of the elementary binors $S_{\alpha\beta},\ldots, S_{\mu\nu}$. In this case the strands are supposed to be drawn so that no portion of a strand passes through α,…,υ – except for the one strand end which terminates there. The general binor is a linear combination with integer (or perhaps rational) coefficients of such outer products, each term having the same set of indices α,…,υ but permuted in some order. Thus, the strand connections for each term are made between the same set of points, but different connections may be made (Fig. 11). In each case it is only the connection scheme which is significant, not the actual paths of the strand.

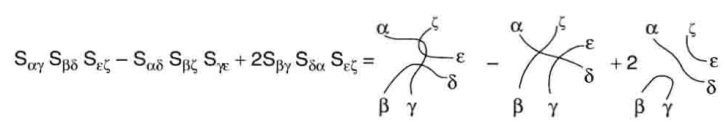


Fig. 11 Two notations for a binor are equated here, the indices in the first case denoting the strand ends in the second case. The actual paths of the strands are not to be significant — only the resulting topological connections are.

Since the strands are not oriented, we have the symmetry

$$S_{\alpha\beta} = S_{\beta\alpha} \tag{3.5}$$

Furthermore the implicit assumptions of commutativity and associativity have been made for outer products:

$$A_{\alpha\ldots\gamma}B_{\delta\ldots\zeta} = B_{\delta\ldots\zeta}A_{\alpha\ldots\gamma} \tag{3.6}$$

$$A_{\alpha\ldots\gamma}(B_{\delta\ldots\zeta}C_{\lambda\ldots\nu}) = (A_{\alpha\ldots\gamma}B_{\delta\ldots\zeta})C_{\lambda\ldots\nu} \tag{3.7}$$

where $A_{\alpha\ldots\gamma}$, $B_{\delta\ldots\zeta}$, $C_{\lambda\ldots\nu}$ stand for outer products of elementary binors. Extending the definition of outer product to all binors by means of the distributive laws.

$$A_{\alpha\ldots\gamma}(B_{\lambda\ldots\nu} \pm C_{\lambda\ldots\nu}) = A_{\alpha\ldots\gamma}B_{\lambda\ldots\nu} \pm A_{\alpha\ldots\gamma}C_{\lambda\ldots\nu}, \tag{3.8}$$

we have 3.6, 3.7, 3.8 holding for all binors with the specified sets of indices. Outer products can occur only between binors with indices in common, whereas sums and differences can occur only between binors whose (unordered) index sets are identical. We must have:

> (3.9) for each unordered set of indices $(\alpha,\ldots,\gamma)$ the binors $A_{\alpha\ldots\gamma}$ form an abelian group with respect to addition.

To these may be added the substitution law:

> (3.10) in any relation between binors the substitution for $\alpha,\beta,\gamma,\delta,\ldots$ of some permutation of $\alpha,\beta,\gamma,\delta,\ldots$ results in an equivalent relation between binors.

A binor resulting from a substitution such as in (3.10) will generally not be equal to the original binor, although it can be if a symmetry is present (cf. (3.5)). Moreover two binors are never equal if the unordered sets of indices are distinct in the two cases (e.g. $S_{\alpha\beta} \neq S_{\alpha\gamma}$).

Finally there is the operation of contraction between binors. The notation is the same as for cartesian tensors: any two index letters may be put equal throughout an expression — or both replaced by some new letter — whereby they become dummy indices and do not contribute to the index structure of the resulting binor. Thus, $A_{\alpha\ldots\gamma\xi\lambda\ldots\nu\xi\rho\ldots\tau} = A_{\alpha\ldots\gamma\theta\lambda\ldots\nu\theta\rho\ldots\tau} = B_{\alpha\ldots\gamma\lambda\ldots\nu\rho\ldots\tau}$ is a contraction of $A_{\alpha\ldots\gamma\xi\lambda\ldots\nu\eta\rho\ldots\tau}$. In any term of a sum, no index may occur more than twice, and if it occurs twice it is a dummy. Any equation is supposed to remain true if any pair of index letters is replaced, throughout the expression by dummies.

In terms of the diagrams, a contraction corresponds to an identification of the two relevant end-points in each outer product diagram. This either joins a pair of strands together or else it produces a closed loop. In the first case the joined pair simply produces a new strand, whence

$$S_{\alpha\beta}S_{\beta\gamma} = S_{\alpha\gamma}. \tag{3.11}$$

In the second case, since a closed loop counts as –2, we have

$$S_{\alpha\alpha} = -2 \tag{3.12}$$

In view of (3.11), we may also think of a contraction between α and β as the drawing of a new strand in the diagram, which joins the point α to β (since $S_{\gamma\alpha}S_{\alpha\beta}S_{\beta\delta} = S_{\gamma\delta}$) Thus we may write

$$A\ldots\zeta\ldots\zeta\ldots = S_{\xi\eta}A\ldots\xi\ldots\eta\ldots \tag{3.13}$$

whenever $A\ldots\xi\ldots\eta\ldots$ is an outer product of elementary binors. Implicit in this is the assumption that (3.6) and (3.7) still remain true (for products of elementary binors) when some of the indices are dummies. Furthermore, extending the definition of contraction linearly, from the case of outer products to any binor, we can maintain (3.13) when $A\ldots\xi\ldots\eta\ldots$ is a general binor. (See Fig. 12 for a diagram illustrating a contraction of a binor.) We also have the distributive property

$$S_{\xi\eta}(A\ldots\xi\ldots\eta\ldots\pm B\ldots\xi\ldots\eta\ldots) = A\ldots\zeta\ldots\zeta\ldots\pm B\ldots\zeta\ldots\zeta\ldots \tag{3.14}$$

Note also the following rule, which is now effectively a special case of (3.13):

$$S_{\xi\eta}A_{\alpha}\ldots\eta\ldots = A_{\alpha}\ldots\xi\ldots \tag{3.15}$$

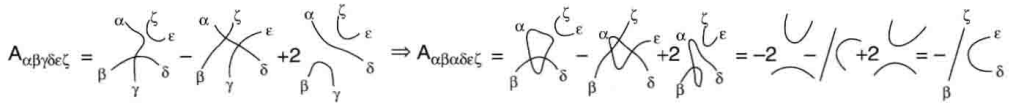


Fig. 12 The binor of Fig. 11 is contracted over α,γ. The final result is $-S_{\beta\xi}S_{\delta\varepsilon}$.

This represents a strand extension from the point η to the point ξ. The usual notations for contracted products will be used, e.g.

$$A\ldots\xi\ldots B\ldots\xi\ldots = S_{\xi\eta}(A\ldots\xi\ldots B\ldots\eta\ldots) \tag{3.16}$$

The commutative, associative and distributive laws hold for contracted products.

It is not the purpose, here, to give a completely formal axiomatic discussion of binor algebra. (Ref. 10) It is hoped, on the other hand, that the foregoing description has been adequate for the present purposes. The essential point about the algebra is that all the binor rules (cf. (3.5), … , (3.16)), with the single exception of (3.12), are just the rules of cartesian tensor algebra where $S_{\alpha\beta}$ plays the part of an n-dimensional Kronecker delta δ_{ab}. The binors correspond to "isotropic" tensors, that is, tensors which are numerically invariant under the full n-dimensional orthogonal group, since these tensors are the ones built up from δ_{ab} (and scalars) by means of outer products, index permutations and sums. (The Levi-Civita symbols are excluded here since they change sign on reflection.) The key difference between the algebra of binors and the algebra of isotropic tensors lies in (3.12): $S_{\alpha\alpha} = -2$. For the contraction of a Kronecker delta we have $\delta_{aa} = n$ where $n > 0$ is the dimension of the space. The binors cannot, in fact, be represented in terms of components in the same manner as are tensors since the dimensionality of the space would have to turn out to be negative (but see section 4 for an alternative representation).

There is one final rule which is convenient (although not absolutely essential) to impose on the binors in order to define the system completely. This is the equation

$$S_{\alpha\beta}\,S_{\gamma\delta} + S_{\alpha\gamma}\,S_{\delta\beta} + S_{\alpha\delta}\,S_{\beta\gamma} = 0 \tag{3.17}$$

The diagram for (3.17) is drawn in fig. 13. (The symbol 0, written without indices, denotes, as is the usual practice, the identity element of the relevant additive abelian group.) We observe that any contraction of the left-hand side of (3.17) vanishes identically by virtue of (3.12). It follows that any expression

Fig. 13 The fundamental binor identity 3.17 in diagrammatic form.

$$(S_{\alpha\beta}\, S_{\gamma\delta} + S_{\alpha\gamma}\, S_{\gamma\beta} + S_{\alpha\delta}\, S_{\beta\gamma})\, A_{\alpha\beta\gamma\delta}$$

must also vanish identically where $A_{\alpha\beta\gamma\delta}$ is a binor, since upon expanding out $A_{\alpha\beta\gamma\delta}$ in terms of elementary binors, we get simply a linear combination of contractions of the bracketed expression, so that each term must vanish identically. Equation (3.17) therefore adds nothing to the theory if it is ultimately the scalars of the theory that are of interest. More important, this indicates that (3.17) will not lead us to a contradiction if we desire to impose it. Equation (3.17) will, in fact, be imposed here as it is useful in reduction formulae. (It should be recalled that isotropic n-dimensional tensors also satisfy certain identities. An outer product of n+1 Kronecker deltas, skew-symmetrized over their first indices must necessarily vanish.)

For any binor $A_{\alpha\ldots\delta}$ we can define a binor norm:

$$\text{norm}\,(A_{\alpha\ldots\delta}) = A_{\alpha\ldots\delta}\, A_{\alpha\ldots\delta}, \tag{3.18}$$

which is a scalar. For the case of a binor representing a spin network the binor norm (3.18) differs by an unimportant non-zero factor from the norm $\|A\ldots\|$ defined in section 2. In fact

$$\|A...\| = \frac{1}{a!b!...d!}\left|\text{norm}\,(A...)\right| \tag{3.19}$$

where a,b,…,d are the spin numbers of the end units of the spin network. (This factor emerges here because of the convention that 1/a! is introduced into the definition of the binor representing a spin network, every time a unit acquires a vertex at both ends.)

In the next section, the following two results will be proved:

$$(-1)^n\, \text{norm}\,(A\ldots) \geq 0 \tag{3.20}$$

where A… has 2n indices and

$$\text{norm}(A\ldots) = 0 \quad \text{if and only if} \quad A\ldots = 0. \tag{3.21}$$

For (3.21) it is essential that (3.17) hold. In fact, if we use the vanishing of the binor norm as a criterion for the vanishing of the binor, it turns out (see section 4) that the only binor identities are those which result from (3.17). If any other identity were to be imposed an inconsistency would result.

In any case, in view of (3.21) we now have an algorithm for deciding whether or not a binor expression represents zero. (A different algorithm will be given in section 5.) Also, combining (3.21), (3.19) and (2.1) we see that:

(3.22) the binor representing a spin network vanishes if and only if the physical situation represented is excluded by non-relativistic quntum mechanics.

In view of (3.22) we will lose no physics from the theory if we think of a spin network as being equated to the binor which represents it.

To conclude this section, some further notation will be introduced, which will extend the utility of the binor diagrams considerably. Define

$$S_{\alpha\beta,\,\gamma\delta} = S_{\alpha\gamma}\,S_{\beta\delta} - S_{\alpha\delta}\,S_{\beta\gamma} \tag{3.22}$$

$$S_{\alpha\beta\gamma,\,\delta\varepsilon\phi} = S_{\alpha\delta}\,S_{\beta\varepsilon}\,S_{\gamma\phi} + S_{\alpha\varepsilon}\,S_{\beta\phi}\,S_{\gamma\delta} + S_{\alpha\phi}\,S_{\beta\delta}\,S_{\gamma\varepsilon} - S_{\alpha\delta}\,S_{\beta\phi}\,S_{\gamma\varepsilon} - S_{\alpha\phi}\,S_{\beta\varepsilon}\,S_{\gamma\delta} - S_{\alpha\varepsilon}\,S_{\beta\delta}\,S_{\gamma\phi} \tag{3.23}$$

and generally

$$S_{\alpha\ldots\delta,\,\kappa\ldots\nu} = \begin{vmatrix} S_{\alpha\kappa} & \ldots & S_{\alpha\nu} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ S_{\delta\kappa} & \ldots & S_{\delta\nu} \end{vmatrix} \tag{3.24}$$

The diagrammatic notation for $S_{\alpha\ldots\delta,\kappa\ldots\nu}$ is r parallel strands with a thicker bar drawn across. (see Fig. 14).

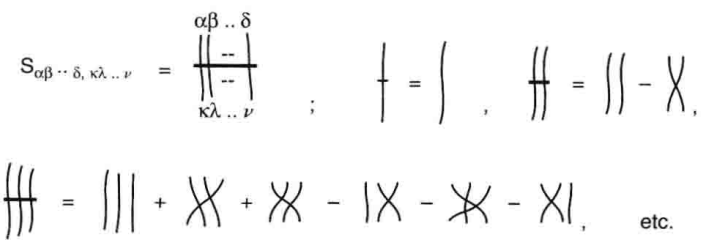


Fig. 14 The binor notation for the expressions (3.22), (3.23), (3.24). Each term in each expansion is an Aitken diagram, so that the sign can be determined as the parity of the number of intersections between strands (drawn according to (3.1) and (3.2)).

Each term is the diagrammatic expansion of $S_{\alpha\ldots\delta,\kappa\ldots\,\nu}$ appears as an Aitken diagram (cf. fig. 7) for the relevant term in the determinant, so the sign of the term is determined by the parity of the number of strand intersections appearing on the plane (assuming (3.1) and (3.2)).

We can represent the binor corresponding to a spin network vertex in terms of these S's. Thus if an a-unit, with strand ends labelled α, .., γ; a b-unit, — labelled δ, .., ζ; and a c-unit, — labelled θ, .., κ, come together at a vertex, the binor for that vertex is

$$\frac{1}{\left\{\frac{1}{2}(a+b-c)\right\}!\left\{\frac{1}{2}(b+c-a)\right\}!\left\{\frac{1}{2}(c+a-b)\right\}!} S_{\alpha..\gamma,\,\lambda..\nu\rho..\tau}\, S_{\delta..\zeta,\,\tau..\rho\phi..\psi}\, S_{\theta..\kappa,\,\psi..\phi\nu..\lambda} \quad (3.25)$$

This should become clearer from the example given in fig. 15. The factorial expression in the denominator removes the redundancy that the ordering of the connecting strands is irrelevant (and see (2.8), (3.4)). The coefficients in (3.25) are thus ±1.

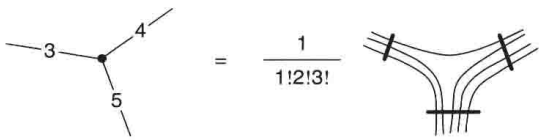


Fig. 15 The binor representing a vertex can be expressed as a contracted product of the binors of fig. 14. The factorials in the denominator remove a redundancy in the expansion.

From the definition of $S_{\alpha..\gamma,\,\delta..\zeta}$ we see that it is skew-symmetrical in $\alpha,\ldots,\gamma$ and in $\delta,\ldots,\zeta$ (and symmetrical between the groups). Hence

$$S_{\alpha..\gamma,\,\delta..\zeta}\, S_{\delta..\zeta,\,\rho..\tau} = r!\, S_{\alpha..\gamma,\,\rho..\tau} \quad (3.26)$$

where $\alpha,\ldots,\gamma$ are r in number. More generally

$$S_{\alpha..\gamma,\,\delta..\zeta}\, S_{\ldots\delta..\zeta\ldots\,,\,\ldots\rho..\tau\ldots} = r!\, S_{..\alpha..\gamma..,..\,\rho..\tau\ldots} \quad (3.27)$$

(see Fig.16). The binors (3.25) corresponding to the various vertices of a spin network may be contracted together to give the binor for the whole spin network, multiplied by r!...t! where r,...,t are the spin numbers of all the internal units in the spin network. By (3.26) and (3.25), this factor is removed if we write only <u>one</u> S..., ... for each unit instead of two (one from each vertex). This is illustrated in Fig. 17.

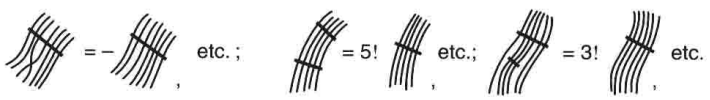


Fig. 16 Some simple properties of S..., ... in diagrammatic form.

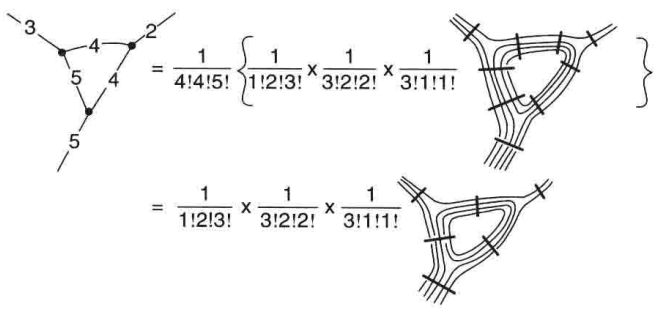


Fig. 17 A simple spin network diagram represented as a binor. The factorials introduced in the denominator corresponding to each internal unit are removed by writing only one S..., ... for each unit.

The other factorials, from (3.25), will remain and it will be convenient to incorporate them into a notation for binors — as has already been done in the case of spin network lines. A line with a non-negative integer n attached denotes, as before, a set of n parallel strands. The integer 1 may be omitted when a single strand is denoted. The line may be omitted altogether if the integer is zero. The thick bar may be used to denote the expression S..., ... and blobs may still be used to denote vertices. We shall use, in addition, the following convention.

(3.28) Any line segment with associated integer n and which is terminated at both ends, either by a vertex blob or by a thick bar, denotes not just n strands but n strands times 1/n!

With this convention, we can remove the factorials from figs. 15, 16 and 17. This is illustrated somewhat more generally in fig. 18.

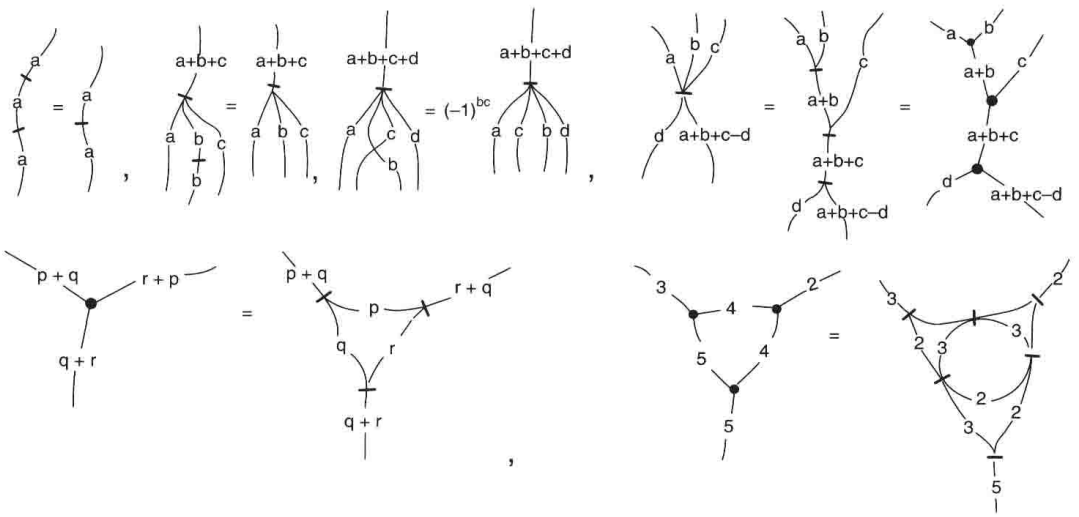


Fig. 18 The notations of figs. 15, 16, 17 can be written more generally and a little more concisely if a set of n strands taken together and denoted by a single line labeled by n. Such a line when terminated at both ends, either by a vertex blob or a thick bar, denotes, instead, 1/n! times x strands. (If n = 1 the label may be omitted.) Henceforth, the right-hand side of the first equation above is the simplest diagram that will be used to denote an isolated a-unit. The fourth equation illustrates how any bar can be written in terms of vertex blobs, whereas the fifth describes the converse operation. The final diagram denotes the strand network for fig. 17. In every case, the total of this strand numbers must be the same at the two sides of each bar.

The fourth equation depicted in fig. 18 illustrates how any bar (i.e. S...,...) in a diagram may be written, instead, in terms of special spin network vertices. The fifth equation shows how the general spin network vertex can be written in terms of the bars (cf. (3.25)). A diagram built up solely of the bars and (multiple) strand lines – i.e. a contracted product of the S...,...'s — will be called a strand network. From the above, we see that strand networks and spin networks are in a sense equivalent, since for any strand network there is a spin network that is equal to it and conversely. However, the operations which suggests themselves are different in the two cases. It will prove very

convenient to be able to work here primarily with the strand networks rather than with the spin networks. The one point to note when drawing strand networks is the following:

(3.29) The sum of the strand numbers for the lines coming from one side of any bar must equal the sum of the strand numbers at the other side of the bar.

This takes the place of the two laws (2.9), (2.10) for spin networks.

One final convention should be adopted in the representation of the spin networks.

(3.30) The binor representing an isolated n-unit is $S_{\alpha..\gamma,\ \delta..\zeta}$, where the strand ends at one end of the unit are labelled $\alpha,..,\ \gamma$ and at the other end, $\zeta,..,\delta$.

Thus, the simplest diagram that should be used for an isolated a-unit is thus right-hand side of the first equation of fig.18. (This is in fact consistent with (2.12), see (4.19).)

4. Proof of Validity of the Binor Calculus A first step in establishing the validity of binors is to set up a kind of "isomorphism" between binor diagrams drawn on a plane and expressions built up from the two-dimensional Kronecker deltas and Levi-Civita symbols: δ^B_A; ε_{AB}, ε^{AB}. (Capital Roman indices take on the values 1, 2. The summation convention is used throughout. The Levi-Civita and Kronecker symbol components are given by $\varepsilon_{AB} = B–A = \varepsilon^{AB}$, $\delta^B_A = 1–(B–A)^2$.) We have

$$\varepsilon_{AB} = -\varepsilon_{BA},\ \varepsilon^{AB} = -\varepsilon^{BA} \tag{4.1}$$

$$\delta^B_A\delta^C_B = \delta^C_A,\quad \delta^B_A\varepsilon_{BC} = \varepsilon_{AC},\quad \delta^B_A\varepsilon^{AC} = \varepsilon^{BC},\quad \varepsilon_{AB}\varepsilon^{AC} = \delta^C_B, \tag{4.2}$$

$$\delta^A_A = 2 \ = \ \varepsilon_{AB}\varepsilon^{AB} \tag{4.3}$$

Also, the identity

$$\varepsilon_{AB}\,\varepsilon_{CD} + \varepsilon_{AD}\,\varepsilon_{BC} + \varepsilon_{AC}\,\varepsilon_{DB} = 0 \tag{4.4}$$

holds, together with the equivalent identities obtained by raising indices of (4.4) with Levi-Civita symbols, e.g.

$$\delta^B_A\delta^D_C - \delta^D_A\delta^B_C = \varepsilon_{AC}\varepsilon^{BD}. \tag{4.5}$$

It is a classical result of invariant theory that every identical relation satisfied by theory ε's and δ's can be built can up from these relations.

A similarity between (4.4) and the binor identity (3.17) is evident — the difference being that ε_{AB} is skew-symmetrical (4.1) whereas $S_{\alpha\beta}$ is symmetrical (3.5). The difference between (4.3) and (3.12) is also one of sign. Therefore if some form of isomorphism between the two systems is to be established, a suitable convention to take care of the sign differences must be employed. This can, in fact, be done very simply in terms of binor diagrams. The diagrams must first be drawn suitably on a plane, which will be thought of as vertical so we can refer to its "top" and "bottom" ends. We must draw the diagram with smooth curves so that, as in (3.2),

(4.6) all intersection points between strands are simple,

In addition,

(4.7) the strands have vertical tangents only at isolated points — at which no intersections must occur;

(4.8) the strands ends are to be extended out vertically at the extreme top and bottom ends of the diagram.

In Fig.19, a diagram is drawn according to these rules.

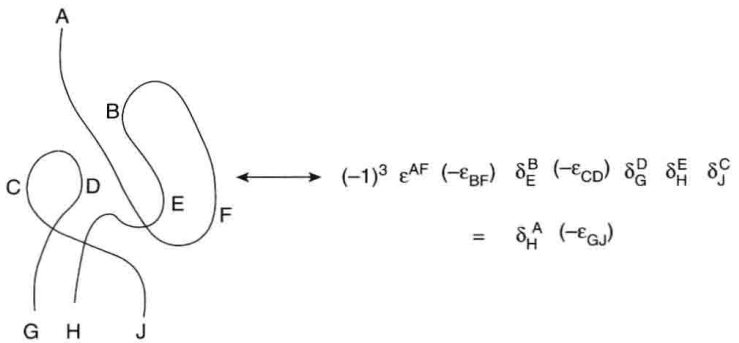


Fig.19 The binor diagram on the left is drawn according to the rules (4.6), (4.7), (4.8). Each point at which a strand is vertical is labelled by a capital letter. The corresponding δ, ε expression has these letters for indices — covariant if the relevant strand portion lies just above the point and contravariant if it lies just below it. A minus sign is introduced for each intersection point in the diagram and for each covariant ε.

Each point at which a strand is vertical is now labelled by a capital Roman letter. These points divide the strands into a number of portions each of which is vertical at no other point. Three essentially distinct situations can now arise with such a strand portion. If P and Q label the two end-points of the portion, we may suppose that Q lies to the right of P. (The situation of P and Q lying vertically one above the other cannot arise for a single such strand line.) Then if P and Q are both lower end-points of the portion, we represent this by $-\varepsilon_{PQ}$. If P and Q are both upper end-points we use $+\varepsilon^{PQ}$. If P is a lower end-point and Q an upper end-point, we use δ^Q_P — whether or not Q lies to the right of P. The appropriately contracted product of all these quantities is formed and then multiplied by $(-1)^r$, where r is the total number of intersection points in the diagram.

The essential virtue of this representation lies in the following two facts.

(4.9) Any continuous deformation of the strand paths with fixed end-points, into another allowable diagram leaves the value of the resulting δ, ε expression unchanged.

(4.10) Any binor identity between diagrams corresponds to a δ, ε identity and vice-versa.

The proof of (4.9) depends on showing that each one of a set of elementary topological deformations — from which the general deformation can be built up — corresponds to a valid δ, ε equation (4.1), (4.2). This is done in fig. 20. As for the proof of (4.10), fig. 21

shows that (3.12) is equivalent to (4.3) and that (3.17) is equivalent to (4.5). The rest is implicit in the relations expressed in fig. 20. (For example, the symmetry of $S_{\alpha\beta}$ is expressed in the skew-symmetry of ε^{AB} by the first equation in fig. 20.) If some of the indices of a valid δ, ε equation are raised or lowered (by ε's), or permuted , this does not affect the validity of the corresponding binor diagram equation. In each case the operation may be achieved by drawing in more lines connected at one end. This cannot alter the validity of a diagram equation. Also, the number of new intersection points introduced has the same parity for each term of a sum, whence the relative signs of the corresponding δ, ε expression remain unaffected.

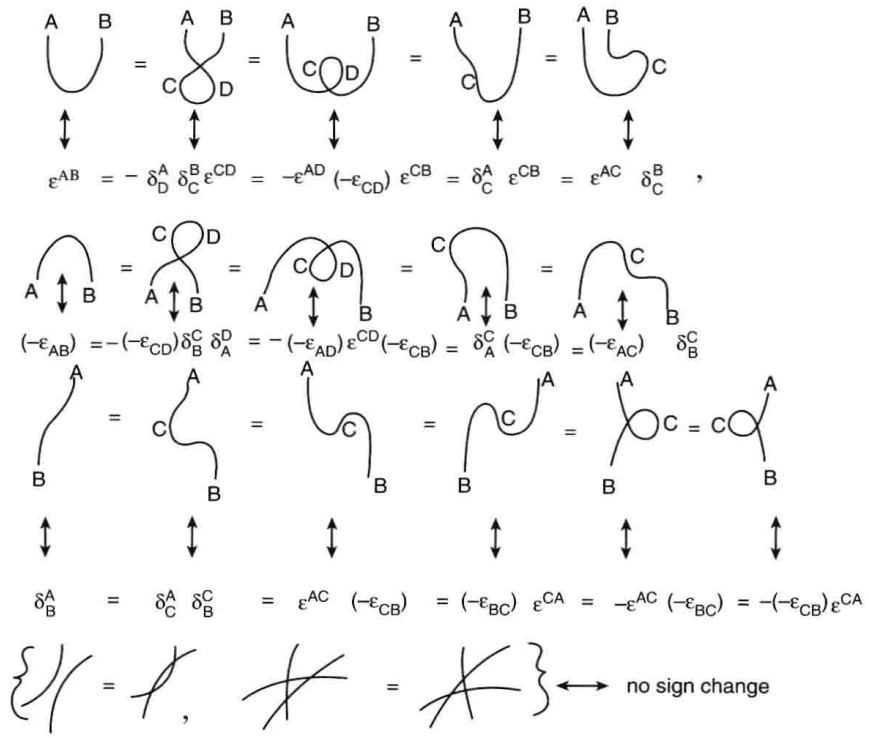

$\varepsilon^{AB} = -\delta^A_D \delta^B_C \varepsilon^{CD} = -\varepsilon^{AD}(-\varepsilon_{CD})\varepsilon^{CB} = \delta^A_C \varepsilon^{CB} = \varepsilon^{AC}\delta^B_C$,
$(-\varepsilon_{AB}) = -(-\varepsilon_{CD})\delta^C_B \delta^D_A = -(-\varepsilon_{AD})\varepsilon^{CD}(-\varepsilon_{CB}) = \delta^C_A(-\varepsilon_{CB}) = (-\varepsilon_{AC})\delta^C_B$
$\delta^A_B = \delta^A_C \delta^C_B = \varepsilon^{AC}(-\varepsilon_{CB}) = (-\varepsilon_{BC})\varepsilon^{CA} = -\varepsilon^{AC}(-\varepsilon_{BC}) = -(-\varepsilon_{CB})\varepsilon^{CA}$
no sign change

Fig. 20 Continuous deformations may be performed on any binor diagram provided the resulting figure satisfies (4.6), (4.7) and (4.8). Although such a deformation may change the form of the corresponding δ, ε expression it will not change the resulting value of the expression (assuming the individual orderings of the upper and of the lower strand ends are unaltered by the deformation).

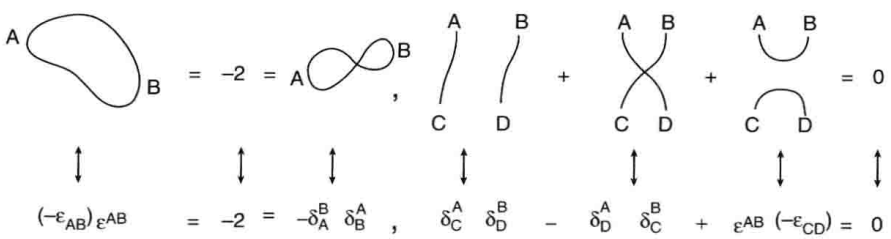

$= -2 =$,
$+ \quad + \quad = 0$
$(-\varepsilon_{AB})\varepsilon^{AB} = -2 = -\delta^B_A \delta^A_B$, $\delta^A_C \delta^B_D - \delta^A_D \delta^B_C + \varepsilon^{AB}(-\varepsilon_{CD}) = 0$

Fig. 21. The two relations which most significantly characterize the binor algebra find more familiar representations in the δ, ε system.

Having thus established our "isomorphism", a number of properties follow at once. Most important of these is:

(4.11) the binor algebra is consistent, although this is, in fact, not hard to see from more direct considerations.

Also, (3.20) and (3.21) are immediate. For let the binor $A_{\lambda..\gamma}$ correspond to the contravariant δ, ε expression $\phi^{L..N}$ and let $\psi_{L...N}$ be the corresponding covariant δ,ε expression (i.e. each $\varepsilon^{..}$ in $\Phi^{L..N}$ is replaced by $-\varepsilon..$ to get $\psi_{L..N}$). Then component by component, we have $\psi_{L..N} = (-1)^r \Phi^{L..N}$ where $\Phi^{L..N}$ has 2r indices. Thus, $(-1)^r \Phi^{L..N} \psi_{L..N} = \sum|\Phi^{L..N}|^2 > 0$ unless $\psi_{L..N} = 0$. That is, $(-1)^r A_{\lambda..\nu} A_{\lambda..\ \nu} > 0$ unless $A_{\lambda..\ \nu} = 0$, as required. Another result, related to this, which follows from the corresponding classical δ, ε result is the fact that no further binor identities, beyond (3.17) can result, except for those built up from (3.17) itself.

Note that the minus sign introduced for each strand intersection results in a curious reversal of the roles of symmetry and skew-symmetry. The diagramatic expansion for $S_{\alpha..\gamma,\ \lambda..\nu}$ in terms of Aitken diagrams given in fig. 14, involves a minus sign each time the number of intersections is odd. Thus, corresponding to $S_{\alpha..\gamma,\ \lambda..\nu}$, the expression $\sigma^{L..N}_{A..C}$ built up out of δ:'s is (apart from factor) a symmetrizing operator rather than a skew-symmetrizing operator:

$$\sigma^{L..N}_{A..C} = \Sigma \{ \delta^{P(L)}_{A} \ldots \delta^{P(N)}_{C} \}, \tag{4.12}$$

the summation extending over all permutations P of L,...,N. On the other hand, since an expression $\chi_{A..C}$ cannot be skew-symmetrical in more than two of its indices unless it is zero, we have the result:

(4.13) any binor which is symmetrical in more than two indices must vanish.

If we consider pairs of binor indices, then symmetry and skew-symmetry revert to their normal roles, since any odd permutation of pairs of elements is still an even permutation of the individual elements. Now, pairs of symmetrical two-dimensional indices may be thought of as constituting a three-dimensional compound index: the set of symmetrical expressions Θ_{AB} (not supposed to be constructed solely from ε's and δ's) is three-dimensional. The role of the Kronecker delta is then taken over by $\frac{1}{2}\sigma^{CD}_{AB}$. It follows from the three-dimensionality that

$$\begin{vmatrix} \sigma^{RS}_{AB} & \sigma^{TU}_{AB} & \sigma^{VW}_{AB} \\ \sigma^{RS}_{CD} & \sigma^{TU}_{CD} & \sigma^{VW}_{CD} \\ \sigma^{RS}_{EF} & \sigma^{TU}_{EF} & \sigma^{VW}_{EF} \end{vmatrix} = \kappa\lambda^{RS,TU,VW}\lambda_{AB,CD,EF}, \quad \begin{vmatrix} \sigma^{RS}_{AB} & \sigma^{TU}_{AB} & \sigma^{VW}_{AB} & \sigma^{XY}_{AB} \\ \sigma^{RS}_{CD} & \sigma^{TU}_{CD} & \sigma^{VW}_{CD} & \sigma^{XY}_{CD} \\ \sigma^{RS}_{EF} & \sigma^{TU}_{EF} & \sigma^{VW}_{EF} & \sigma^{XY}_{EF} \\ \sigma^{RS}_{GH} & \sigma^{TU}_{GH} & \sigma^{VW}_{GH} & \sigma^{XY}_{GH} \end{vmatrix} = 0 \tag{4.14}$$

where $\lambda^{RS,TU,VW}$ and $\lambda_{AB,CD,EF}$ are effectively the three-dimensional Levi-Civita symbols, being defined up to proportionality by the skew-symmetry in their pairs of indices, κ being a convenient constant.

The binor equivalents of (4.14) must also hold, where the expressions now have pairs of skew-symmetrical indices:

$$\begin{vmatrix} S_{\alpha\beta,\rho\sigma} & S_{\alpha\beta,\tau\upsilon} & S_{\alpha\beta,\phi\chi} \\ S_{\gamma\delta,\rho\sigma} & S_{\gamma\delta,\tau\upsilon} & S_{\gamma\delta,\phi\chi} \\ S_{\varepsilon\zeta,\rho\sigma} & S_{\varepsilon\zeta,\tau\upsilon} & S_{\varepsilon\zeta,\phi\chi} \end{vmatrix} = \kappa L_{\rho\sigma,\tau\upsilon,\phi\chi} L_{\alpha\beta,\gamma\delta,\varepsilon\zeta}, \quad \begin{vmatrix} S_{\alpha\beta,\rho\sigma} & S_{\alpha\beta,\tau\upsilon} & S_{\alpha\beta,\phi\chi} & S_{\alpha\beta,\psi\omega} \\ S_{\gamma\delta,\rho\sigma} & S_{\gamma\delta,\tau\upsilon} & S_{\gamma\delta,\phi\chi} & S_{\gamma\delta,\psi\omega} \\ S_{\varepsilon\zeta,\rho\sigma} & S_{\varepsilon\zeta,\tau\upsilon} & S_{\varepsilon\zeta,\phi\chi} & S_{\varepsilon\zeta,\psi\omega} \\ S_{\eta\theta,\rho\sigma} & S_{\eta\theta,\tau\upsilon} & S_{\eta\theta,\phi\chi} & S_{\eta\theta,\psi\omega} \end{vmatrix} = 0 \qquad (4.15)$$

It follows from (2.11) that the binor representing the spin network vertex at which three 2-units come together, is skew-symmetrical in its three pairs of skew-symmetrical indices. Thus, from (3.25), we can put

$$L_{\alpha\beta,\gamma\delta,\varepsilon\zeta} = S_{\alpha\beta,\mu\nu} S_{\gamma\delta,\nu\xi} S_{\varepsilon\zeta,\xi\mu} \qquad (4.16)$$

and it, from which it then in fact, turns out that $\kappa = -1/4$. These relations will be needed later.

More generally, we can consider n-tuples of symmetrical capital indices and, equivalently, n-tuples of skew-symmetrical binor indices. The role of the Kronecker delta is taken over, respectively, by

$$\frac{1}{n!}\sigma^{R\ldots T}_{A\ldots D} \quad \text{and} \quad \frac{1}{n!} S_{\alpha\cdot\cdot\delta,\rho\cdot\cdot\tau} \qquad (4.17)$$

in the two cases (A,..,D; α,..,δ are each n in number). Since the symmetrical expressions $\Theta_{A\ldots D}$ (not intended as constructed solely from ε's and δ's) form an (n+1)–dimensional vector space, it follows that the trace of the corresponding "Kronecker delta" must be n+1, whence

$$\sigma^{A\ldots D}_{A\ldots D} = (n+1)\, n! = (n+1)! \qquad (4.18)$$

Thus

$$S_{\alpha\cdot\cdot\delta,\alpha\cdot\cdot\delta} = (-1)^n (n+1)! \qquad (4.19)$$

(The contraction can be performed diagrammatically with no additional intersection lines if n ε's and n $(-\varepsilon..)$'s are used. Hence the factor $(-1)^n$.) An alternative derivation of (4.19) is given in section 5. The convention (2.12) is now seen to be consistent with (3.26) and (3.30).

It will be necessary, in section 6, to be able to treat approximate binor equations. The notation for binors

$$A_{\alpha..\gamma} \sim B_{\alpha\ldots\gamma} \qquad (4.20)$$

will be used to denote

$$|\text{norm}\,(A_{\alpha..\gamma} - B_{\alpha..\gamma})| << |\text{norm}\,(A_{\alpha..\gamma})| \qquad (4.21)$$

(where "<<" means that the ratio of the left-hand to the right-hand expression is very small).

Let the δ, ε expressions $\alpha_{A..C}$, $\beta_{A..C}$ correspond to $A_{\alpha..\gamma}$, $B_{\alpha..\gamma}$, respectively. In the discussion given earlier, which was used to prove (3.20), it emerged that |norm (A...)| was simply the sum of the squares of the components of α.... Thus (4.21) states that the "vectors" $\alpha_{A..C}$ and $\beta_{A..C}$ in the appropriate 2^r dimensional space are very nearly equal in direction and have magnitudes whose ratio is very nearly unity. It follows that (4.20) is symmetric:

$$A\ldots \sim B\ldots \Rightarrow B\ldots \sim A\ldots ; \tag{4.22}$$

transitive:

$$A\ldots \sim B\ldots,\quad B\ldots \sim C\ldots \Rightarrow A\ldots \sim C\ldots ; \tag{4.23}$$

and appropriately additive:

$$A\ldots \sim B\ldots ,\quad C\ldots \sim D\ldots ,\quad A\ldots \not\sim -C\ldots \Rightarrow A\ldots + C\ldots \sim B\ldots + D\ldots . \tag{4.24}$$

Now suppose the binor $P_{\alpha..\gamma,\lambda..\nu}$ (with the same number of indices in each of its two groups) is such that for any binor $X_{\alpha..\gamma}$, we have

$$|\text{norm}(P_{\alpha..\gamma,\lambda..\nu}X_{\lambda..\nu})| \leq n|\text{norm}(X_{\alpha..\gamma})| \tag{4.25}$$

where n is fixed independently of X... and is chosen to be not large compared with unity. (In fact n = 2 will do.) Then[†]

$$P_{\alpha..\gamma,\,\lambda..\nu}A_{\lambda..\nu} \sim A_{\alpha..\gamma} \sim Q_{\alpha..\gamma,\,\lambda..\nu}A_{\lambda..\nu} \Rightarrow P_{\alpha..\gamma,\,\lambda..\nu}Q_{\lambda..\nu,\rho..\tau}A_{\rho..\tau} \sim A_{\alpha..\gamma}. \tag{4.26}$$

To prove this, assume the left-hand side of the implication holds and substitute $Q_{\lambda..\nu,\rho..\tau}A_{\rho..\tau} - A_{\lambda..\nu}$ for $X_{\lambda..\nu}$ in (4.25). Then[†]

$$\begin{aligned}|\text{norm}(P_{\alpha..\gamma,\lambda..\nu}\,Q_{\lambda..\nu,\rho..\tau}\,A_{\rho..\tau} - P_{\alpha..\gamma,\lambda..\nu}\,A_{\lambda..\nu})| &\leq n\ |\text{norm}(Q_{\lambda..\nu,\rho..\tau}\,A_{\rho..\tau} - A_{\lambda..\nu})| \\ &\ll |\text{norm}(A_{\lambda..\nu})|\end{aligned}$$

and also

$$|\text{norm}(P_{\alpha..\gamma,\lambda..\nu}\,A_{\lambda..\nu} - A_{\alpha..\gamma})| \ll |\text{norm}(A_{\lambda..\nu})|.$$

Combining these two results we get the right-hand side of (4.26), as required.

Another implication of the binor to δ, ε correspondence is Schwarz' inequality for binors:

$$A_{\alpha..\gamma}A_{\alpha..\gamma}B_{\lambda..\nu}B_{\lambda..\nu} \geq A_{\alpha..\gamma}B_{\alpha..\gamma}A_{\lambda..\nu}B_{\lambda..\nu}. \tag{4.27}$$

Both sides are positive (see (3.20)), A... and B... having the same number of indices. Equality in (4.27) implies A... and B... are proportional. Thus approximate equality implies approximate proportionality.

$$\text{norm}(A\ldots)\ \text{norm}(B\ldots) \sim (A_{\alpha..\gamma}B_{\alpha..\gamma})^2 \Rightarrow A_{\alpha..\gamma} \sim \kappa B_{\alpha..\gamma} \tag{4.28}$$

for some κ. (To see this, we may argue from continuity, after first normalizing A... and B..., if desired.)

[†] Note that $A_{\lambda..\nu} \sim B_{\lambda..\nu} \not\Rightarrow P_{\alpha..\,\gamma,\,\lambda..\nu}A_{\lambda..\nu} \sim P_{\alpha..\,\gamma,\,\lambda..\nu}B_{\lambda..\nu}$.

Finally, the relevance of the binor calculus to physics must be established, namely (2.1) and the more general probability law (2.2). Consider a spin network representing some physical situation. It will be convenient to draw the spin network and the corresponding binor on a plane according to the conventions of (4.6), (4.7) and (4.8). (Of course this implies no such restriction on the actual physical situation represented!) The time-ordering of events is conceived of as proceeding from the bottom of the diagram to the top. Thus, the incoming end units are those depicted as entering at the bottom of the diagram and the outgoing end units, those depicted leaving at the top. It will simplify considerations here if, moreover, we can assume that all the incoming units have spin number zero. Then all the strand ends occur at the top of the diagram. In fact this does not affect the generality of the situations considered since each incoming unit with non-zero spin number s can be considered as having originated from a O-unit as one of a pair of s-units, the other of which plays no further part in the interaction scheme. In terms of diagrams, since o-unit lines can be ignored, this has the effect that free ends are turned away from the bottom of a diagram and they terminate instead at the top.

The initial wave function corresponds to zero angular momentum and can therefore be represented by a scalar ψ. Suppose, next, two s-units are produced out of this zero angular momentum state. Each s-unit individually would have a wave function which could be represented as a symmetric s-index spinor, but we must consider the two together here. Thus the wave function for the two s-units together can be represented as a spinor $\psi^{A..C,D..F}$ which is symmetrical in each of its two sets of s indices. However $\psi^{A..C,D..F}$ represents a state with zero total angular momentum and it therefore still amounts effectively to a scalar. Hence

$$\psi^{A..C,D..F} = \Phi\varepsilon^{AR}..\varepsilon^{CT}\sigma^{D..F}_{R..T} \tag{4.29}$$

where the $\sigma^{\cdots}_{\cdots}$ factor (cf. (4.12)) ensures symmetry in both A,...,C and D,...,F. We can incorporate the introduction of further pairs of r-units, t-units, etc. by extending (4.29) to include further factors $\varepsilon^{KU}..\varepsilon^{MW}\sigma^{N...P}_{U...W}$, etc., one for each pair of units introduced. Thus we have a wave function $\psi^{A..C,..,K..M,..}$ which has a group of s symmetrical spinor indices for each s-unit which is present at this stage and which, apart from a scalar multiplier is the δ, ε expression corresponding to the (rather trivial) spin network so far (see example in fig. 22).

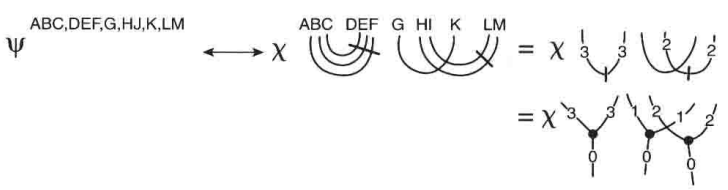


Fig. 22 The required "incoming" units (here a 3-unit, a 1-unit and a 2-unit) are considered to have arisen initially from o-units. The wave function $\psi\cdots$ is then a scalar χ, times the δ, ε expression which corresponds to the spin network, or binor diagram, representing the production of these units.

Next, suppose that an r-unit and an s-unit combine together to form a t-unit. Let the symmetrical sets of indices D..J and P..U in $\psi^{\cdots,D..J,\ldots,P..U,\cdots}$ correspond, respectively, to

this r-unit and this s-unit. The production of the t-unit corresponds to the reduction of the spinor — with respect to these two sets of indices only — to one which, for fixed values of the remaining indices, is irreducible corresponding to the spin value t. $\frac{1}{2}\hbar$. Thus, the new wave function is proportional to

$$\psi^{\ldots,D..FG..J,\ldots,P..RS..U,\ldots}\,\varepsilon_{GP}\ldots\varepsilon_{JR}\,\sigma^{X\ldots Z}_{D..FS..U} \tag{4.30}$$

where the symmetrical set X…Z corresponds to the new t-unit. The corresponding similar right-hand expression is then contracted on, for any other pair of units which combine together. Conversely, whenever an r-unit splits up into two other units, the process is just the reverse of (4.30). Thus, if the symmetrical set D…J corresponds to the r-unit, the new wave function is proportional to

$$\psi^{\ldots,D\ldots FG\ldots J,\ldots}\,\varepsilon^{NQ}\ldots\varepsilon^{PS}\,\sigma^{U\ldots\ldots W}_{D\ldots FN\ldots P}\;\sigma^{X\ldots\ldots Z}_{G\ldots JQ\ldots S} \tag{4.31}$$

where the symmetrical sets U…W and X…Z correspond to the two new units appearing. This process is illustrated in fig. 23. The final wave function is seen to be, apart from a non-zero scalar multiplier, simply the δ, ε expression corresponding to the binor of the relevant spin network. It follows that this resulting binor vanishes if and only if the corresponding physical situation has zero probability. This establishes (3.22) and hence also (2.1).

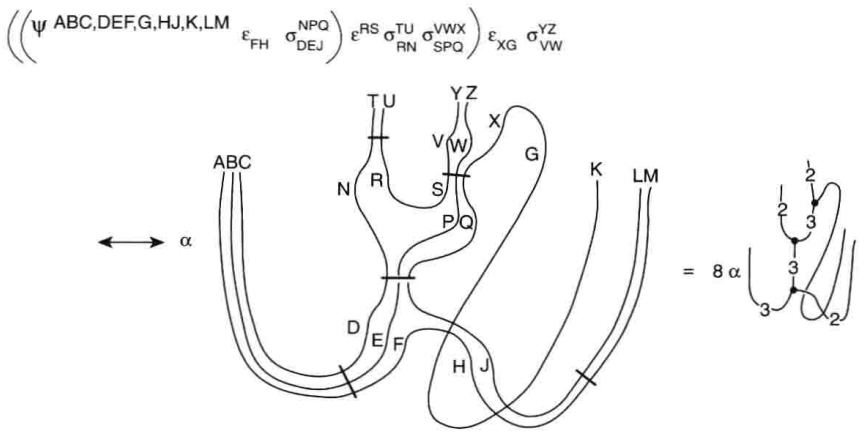


Fig. 23 To the initial wave function given in fig. 22 is applied a succession of three operators. The first represents the combining of a 3-unit and a 2-unit to form a 3-unit, the second represents a subsequent splitting of this 3-unit into a 2-unit and another 3-unit which combines in the final operation with a 1-unit to form a 2-unit. The final wave function is proportional to the δ, ε expression which corresponds to the relevant spin network.

In order to establish the more general probability law (2.2) some simple properties of spin network binors must be noted. First, the binor of any spin network which has just two end units — an r-unit and an s-unit — must necessarily vanish unless r = s. This is

conservation of total angular momentum, but it is also easy enough to see directly from the expansion of the S...,...'s for the binor. (If r > s, and only the S...,... belonging to the end r-unit is not expanded out, then it follows that every term involves S...,...α..α = 0.) If r = s, then similar reasoning shows that the binor must be a multiple of S...,... — the binor of an isolated r-unit. Now, consider any spin network ξ which is a tree, with end units with respective spin numbers a,b,..,d,x. Consider, also, a similar spin network η, which has the same network structure as ξ and corresponding end units with spin numbers a,b,...,d,y. Join ξ to η at the corresponding a-unit, b-unit,...,d-unit leaving only the x-unit and the y-unit as end units. Then it follows from the above that the resulting binor must vanish unless ξ and η are in fact, identical (see fig. 24), in which case it is, say, θ times the binor of an isolated x-unit. Consider, next, the spin network Φ which is obtained from two copies of ξ by joining along the x-unit only. Denote the corresponding binor by $F_{\alpha\ldots\delta,\kappa\ldots\nu}$ where α,..,δ correspond to one set of end a,b,..,d-units κ,...,ν to the other set taken in the corresponding order. From the above considerations it follows that

$$F_{\alpha\ldots\delta,\kappa\ldots\nu}F_{\kappa\ldots\nu,\pi\ldots\tau} = \zeta F_{\alpha\ldots\delta,\pi..\tau} \tag{4.32}$$

where

$$\zeta = x!a!b!...d!\theta \tag{4.33}$$

the factorials arising because of convention (3.28). We can evaluate θ by contracting the left-hand expression in fig. 24.

the result being

$$\theta = \frac{\|\xi\|}{x+1} \tag{4.34}$$

(see (2.12)). (The explicit evaluation of ||ξ|| can, in fact, be easily achieved; see section 5.)

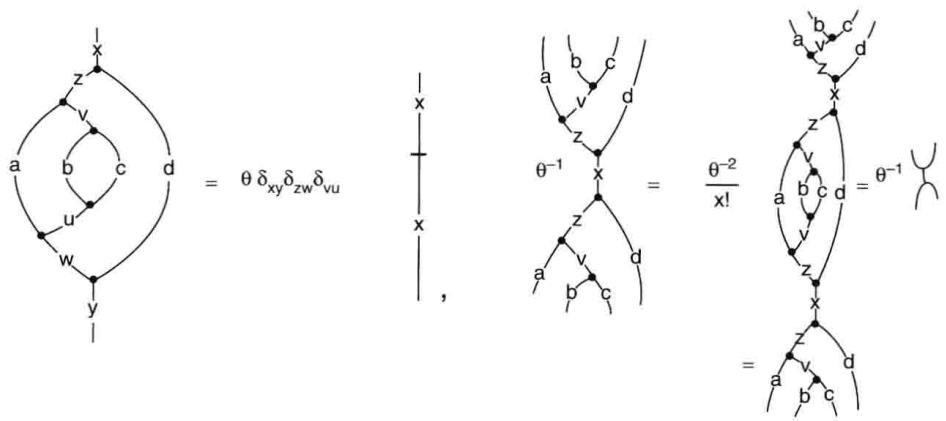


Fig. 24 The equation on the left expresses conservation of total angular momentum in a simple case. The right-hand expression is effectively an idempotent.

From (4.32), we see that $\zeta^{-1}F_{\alpha..\delta,\kappa..\nu}$ is an idempotent. Also, the various corresponding idempotents obtained by changing the internal spin number values must all be orthogonal to each other because of the above remarks. Denoting the binor of the known spin network κ (see section 2; fig 4) by $K_{\alpha\ldots\delta,\rho..\tau}$ where $\rho,..,\tau$ correspond to end units not connected to ξ, it follows that the probability of the spin number values given by ξ is Ref. 10.

$$\frac{|K_{\alpha..\delta;\rho..\tau}\,(\zeta^{-1}F_{\alpha..\delta,\kappa..\upsilon})K_{\kappa..\upsilon;\rho..\tau}|}{|K_{\alpha..\delta;\rho..\tau}\,K_{\alpha..\delta;\rho..\tau}\,|} \tag{4.35}$$

Now, $|K_{\alpha..\delta;\rho..\tau}\,F_{\alpha\ldots\delta,\kappa\ldots\nu}\,K_{\kappa..\nu;\rho..\tau}| = a!b!...d!|\,\mathrm{norm}(G_{\alpha..\delta;\rho..\tau})|$, where $G_{\alpha\ldots\delta;\rho..\tau}$ is the binor of the spin network ω (see section 2; fig. 4) which is the known network κ with the unknown "tree" part ξ attached. The result (2.2) now follows from (4.35) by (4.34), (3.19), (3.18).

ADDENDUM, WRITTEN IN APRIL 2010

The manuscript breaks off here in an incomplete state. If my memory serves me, I had embarked upon—or perhaps resumed work on—this manuscript while staying in Seattle, Washington, USA, for a couple of weeks in August 1967, following the *Battelle Rencontres* meeting of mathematicians and physicists organized by John A. Wheeler and Cécile M. DeWitt (see Chapter 36, Volume 2, pp.62–176 of these *Collected Works*). Although I had seemed to be getting within sight of completing this article during that period, it had become increasingly clear to me that I needed to break off the writing of it in order that I could embark on the apparently more urgent task of writing up my 12-lecture course for this meeting, which subsequently resulted in Chapter 36, in Volume 2 of these *Collected Works*. As things turned out, I did not find a suitable opportunity to return to the manuscript under consideration here, as other matters continued to intervene, distracting me from completing this work. This remains the case, even to this day; yet the route whereby I had intended to achieve what is needed for completion still remains fairly clear to me, so I can at least provide a very brief outline of the intended argument here.

The main result that needs to be shown is what has become known as the *geometry theorem* for spin networks. Suppose we have a spin network possessing several end units with a very large spin numbers—these being referred to as *large* end units. We can contemplate the idea that each such large end unit might be interpreted as describing a spinning body whose axis of spin provides us with a classical-like picture of a *direction* within some abstract Euclidean space. This Euclidean space would not be present initially as a background with respect to which the spins would be defined, since no such background has been introduced, the combinatorial spin-network being all that we have been presented with. Instead, this Euclidean 3-geometry would have to *emerge*, in some way from the structure of the given spin network itself. This might come about initially only as an approximation, but we could imagine that when the structure of the spin network becomes larger and more complicated, or *firmer* in some appropriate sense, then the 3-geometry would become better defined.

In fact, it is really the geometry of *directions* in Euclidean 3-space that we aim to find, since directions are what classical spin axes provide. An appropriate idea for obtaining this geometry is described in the Appendix to my Adams Prize essay, which is Chapter 28 of this Volume, pp.579–729, the Appendix being pp.714–718.

The procedure for defining an 'angle' between large end units of respective spin numbers M and N is described on p.717, with particular reference to the figs. A6–A9 appearing there. Fig. A6 shows the initial situation. In Fig. A7, a 1-unit is taken from the M-unit to the N-unit, and we suppose that it is measured to have reduced the M-unit to an (M–1)-unit and we examine the respective probabilities p or 1–p, for a measurement to find the N-value to be increased to N+1 or reduced to N–1, as calculated by the procedures of the present article for the given spin network. Then a first estimate of the angle θ between the two spin axes would be given by

$$\cos\theta = 2p - 1.$$

However, we need to repeat the procedure in order to ascertain whether this probability is to be considered to be a purely *quantum* probability arising from an angle between well-defined spin directions, or else arising from some other 'knowledge uncertainty' inherent in the arrangement. The criterion adopted is that this probability is to be taken as being indeed a quantum probability resulting from such an angle if the calculated probability in the second operation of this procedure, as illustrated in fig. A8, is not substantially affected by the result of the first measurement (i.e. that the additional knowledge obtained from the first measurement does not influence the probability). We then find that the spin network of Fig. A9 differs by only a small amount from ½cosθ times the original spin network of fig. A6. When this happens, we say that the large end units under consideration have a *well-defined angle* θ *between their spin axes*.

We note the use of vague words 'large', for the size of the spin values of the relevant units and 'small' for the slight changes in the probabilities that may occur between the first and second spin measurement of the N-unit and for the spin networks of figs. A6 and A9. These refer to the fact that we are not to expect the small differences between these probabilities to be exactly zero, or that the relevant spin networks should agree exactly, except in some limit where the spin numbers become infinite. Accordingly, we relate these measures of 'small' and 'large' to the corresponding notions involving the symbols '~' and '«' that occur in (4.20) and (4.21) and later in Section 4 of this article. These approximate notions have to be made more precise in terms of some small parameter Δ and its large counterpart Δ^{-1}. Making these notions precise requires a bit of organization, but the essential relations and inequalities are to be found in Section 4.

What needs to be shown is that if we have a spin network with several large end units with well-defined angles between all of them, then, first, this fact does not significantly depend on the order in which the spin measurements are made and, second, the resulting 'θ-angles' obtained by the above procedure are consistent, in the limit $\Delta \to 0$, with angles between directions in Euclidean 3-space. For this consistency, one needs to show two things. First, there is an identity between the cosines of all the angles that arise between any four different directions in Euclidean 3-dimensional space, and this has to be closely satisfied by the calculated θ-angles between any four different large end units of the spin network. This identity is the vanishing of a 4×4 determinant, which amounts to the right-hand

determinant (4.14), whose vanishing has indeed been established. Second, there are some inequalities that need to be satisfied, which come from the positive-definite nature of the 3×3 matrix on the left in (4.14) (ensuring that the metric of our space has a positive-definite, rather than Lorentzian spatial geometry).

One further matter must be established, in order to complete the argument. This is to confirm that the spin-network geometry is indeed 3-dimensional rather than merely giving directions in a 2-dimensional subspace of Euclidean 3-space. For this, specific examples may be presented, for which the θ-angles are not consistent with geometry in a 2-dimensional plan.

Clearly there are many details to be filled out in this argument, but the essentials are as I have described them. It may also be pointed out that a number of other significant and curious properties of spin networks can be found in the articles in Chapters 45 and 46 in Volume 2 of these *Collected Works*. Mention should also made of the more recent applications of a somewhat different version of spin-network theory in the loop-variable approach to quantum gravity (see Ashtekar, A. and Lewandowski, J. (2004) *Background independent quantum gravity: a status report.* [gr-qc/0404018]).

REFERENCES

1. E.T. Bell, *Men of Mathematics*, Fireside Books (Holiday House), 1937.
2. H. Weyl, *Das Kontinuum*, Chelsea, New York, 1956; A. Robinson, *Non-Standard Analysis*, Princeton University Press, Princeton, NJ, 1966.
3. E. Schrödinger, *Science and Humanism*, Cambridge University Press, Cambridge, 1956.
4. P. Dirac, *The Principles of Quantum Mechanics*, Clarendon Press, Oxford, 1930.
5. A.P. Yutsis, I.B. Levinson, and V.V. Vanagas (1962) *Mathematical Apparatus of the Theory of Angular Momentum* (originally in Russian; translated into English for the National Science Foundation, Washington, and IPST, Jerusalem); R.J. Ord-Smith, *Phys. Rev.* **94**, 1227 (1954); A.R, Edmonds, *Angular Momentum in Quantum Mechanics*, Princeton University Press, Princeton, NJ, 1957.
6. A.R. Edmonds, *Angular Momentum in Quantum Mechanics*, Princeton University Press, Princeton, NJ, 1957.
7. A.C. Aitkin, *Determinants and Matrices*, Oliver & Boyd, Edinburgh, 1958.
8. *See* Chapter 45, R. Penrose, Applications of negative dimensional tensors. *Combinatorial Mathematics and its Applications* (Proc. Conf., Oxford, 1969), ed. D. J. A. Welsh, pp. 221–44, Academic Press, London, 1971. [post hoc]
9. *See* 8.
10. J. von Neumann, *Mathematical Foundations of Quantum Mechanics,* Princeton University Press, Princeton, NJ, 1955.

32
ON THE STRUCTURE OF CAUSAL SPACES

Proc. Camb. Phil. Soc. **63** (1967), 481–501.
With E. H. Kronheimer

In recent years, causal spaces (now usually called causal sets) have interested researchers working in quantum gravity. For a contemporary assessment see, for example, Fay Dowker, 2008, Causal sets and the deep structure of spacetime (arXiv:gr-qc/0508109v1 26 Aug 2005).

On the structure of causal spaces

By E. H. KRONHEIMER and R. PENROSE
Birkbeck College, London

(*Received* 3 *June* 1966)

Abstract. The paper examines the structure obtained by abstracting from the conventional (manifold) representation of relativistic space-time the concept of an event-set equipped with two partial orderings, whose counterparts are the notions 'causally precedes' and 'chronologically precedes in the history of some observer'.

0·1. *Introduction.* The usual pseudo-Riemannian manifolds studied as models of space-time in the general theory of relativity admit a relation of causality between the points (or 'events') of the manifold: an event x *causally precedes* an event y if the interpretative principles would allow an occurrence at x to influence what happens at y, or if a message could be transmitted from x to y. At the same time such models contain a second, more restrictive relation of *chronological precedence*, corresponding to the possible time-ordering of events on the world-line of an idealized observer, whose velocity is less than that of light. In certain manifolds such an analysis of causal and chronological precedence may become ambiguous or impossible; but we shall treat the existence of consistent relations of this character as a criterion of admissibility.

We shall discuss the notion of causal and chronological precedence on an axiomatic basis, describing the event-space as a set equipped with two partial orderings: the *causality* $\prec$, and the *chronology* $\ll$. The conditions these must satisfy are exhibited in section 1 (axioms I to VI): in particular $x \prec y$ ('x causally precedes y') whenever $x \ll y$ ('x chronologically precedes y'). In addition we write '$x \to y$' if $x \prec y$ but not $x \ll y$: the relation $\to$ will be called the *horismos*. The set so equipped is a *causal space*, and $\prec$, $\ll$, $\to$ are its *causal relations*. (In the special theory of relativity such relations have been studied by Zeeman (13).)

The model of space-time as a manifold motivates the choice of results and definitions, but there is no attempt to introduce sufficient axioms to reproduce all relevant properties of that model. Indeed the axiom system will be kept as small and 'physically reasonable' as possible. This has two aims:

(*A*) To analyse unusual or pathological features of space-time manifolds for which no criterion of physical inadmissibility may otherwise be evident. (This type of causal analysis of the global structure of manifolds may prove useful in the context of recent work exemplified by (8,9,5). Other kinds of pathological space-time manifold are discussed in, for example, (4, 7).)

(*B*) To admit structures which can be very different from a manifold. The possibility arises, for example, of a locally countable or discrete event-space equipped

with causal relations macroscopically similar to those of a space-time continuum. (Discrete models of space-time have been suggested by a number of authors as a possible way of avoiding the infinities of quantum field theory, etc. See, for instance, **(12, 11, 3, 6, 2, 1)**.)

0·2. *Manifolds as causal spaces.* We shall not give proofs of the causal properties of space-time manifolds (which are referred to only parenthetically in sections 1 to 5). It may, however, clarify matters to begin by identifying the *natural causal structure* associated with an admissible manifold.

This is, of course, simplest in the case of n-dimensional Minkowski space-time, L^n. ($n \geqslant 2$. It will prove convenient, throughout, not to restrict attention to 4-dimensional manifolds.) Defining L^n as the set of real n-tuples $x = (x_0, \ldots, x_{n-1})$, equipped with the form

$$Q(x,y) = (x_0 - y_0)^2 - \sum_1^{n-1} (x_i - y_i)^2,$$

we write '$x \to y$', '$x \ll y$', or '$x \prec y$' to mean that $x_0 \leqslant y_0$ and that $Q(x,y)$ is, respectively, zero, strictly positive, or not strictly negative; or equivalently (with the usual definition of null-cones and the words 'inside' and 'outside') that y lies respectively on, inside, or not outside the future null-cone of x; or yet again that x lies respectively on, inside, or not outside the past null-cone of y. As a causal space, L^n is paradigmatic: it satisfies every condition we shall discuss in this paper.

Turning to the general case, we have to consider an n-dimensional differentiable manifold X equipped with a pseudo-Riemannian metric g of signature $2-n$: a metric whose only function for our purpose is to identify a null-cone N_x in the tangent-space T_x at each point x of X and enable any vector t in T_x to be identified as *null*, *timelike*, or *spacelike* according as it belongs to N_x, to its interior, or to its exterior—that is, according as $g(t,t)$ equals, exceeds, or is less than zero. The removal of the zero vector and exterior of N_x decomposes T_x into two separated sets, whose members are to be labelled '*future-directed*' and '*past-directed*'. The properties of the metric ensure that this time-labelling can be carried out in a continuous manner locally; but whether it can be achieved continuously over the whole manifold depends on the global connectivity properties in relation to the null-cones, and we make the possibility of doing so a *first restriction* on the admissibility of the manifold.

Let us agree to call a differentiable parametrized curve in X whose (oriented) tangents are all timelike (resp. all non-spacelike) a *timelike* (resp. *non-spacelike*) *path*. Our first restriction on X enables us to define an *orientation* for any non-spacelike path: calling it *positive* if its tangent vectors are everywhere future-directed. Let us next write 'xTy' (resp. 'xNy') to mean that there exists a positively oriented timelike (resp. non-spacelike) path Γ and real numbers s, t such that $x = \Gamma(s)$, $y = \Gamma(t)$, and $s < t$. These relations satisfy the following proposition:

For each x, y, z in X:

(*a*) *if xTy, then xNy;*
(*b*) *if xNy and yNz, then xNz;*
(*c*) *if xTy and yNz, or if xNy and yTz, then xTz.*

(0·2·1)

Now we define the three causal relations on X as follows:

$x \prec y$ *if and only if either* xNy *or* $x = y$;

$x \ll y$ *if and only if* xTy;

$x \to y$ *if and only if either* xNy *and not* xTy *or* $x = y$.

The proposition (0·2·1) will ensure that these relations satisfy the axioms for a causal structure on X, provided that we impose a *second restriction* on X, amounting to a prohibition on either observers or signals travelling into their own pasts. To be precise, we exclude closed paths by the requirement that xNx for no point x in an admissible manifold.

To illustrate: the most obvious metrics on a Moebius band violate either the first or the second restriction; but Example 6–4 exhibits an admissible metric. Orientability of the manifold is not required. On the other hand closed—i.e. compact—manifolds are excluded: see 1·5. (This result is well-known (**14**).)

It might reasonably be thought that, as a causal space, a manifold must be, in some sense, 'locally Minkowskian'. More precisely: if we define a *causal isomorphism* to be a one-to-one mapping f of one causal space onto another such that both f and f^{-1} preserve causal relations, we might consider the following proposition:

For each point x of the n-dimensional manifold X we can find an arbitrarily small open neighbourhood U of x, together with a homeomorphism h of U onto an open subset V of L^n, *such that h is a causal isomorphism.* (0·2·2)

As it stands (0·2·2) leaves unspecified the causal structures with which U and V are equipped. There are two 'natural' choices: (*a*) we equip U with the natural causal structure of X *restricted* to the subset U (see 1·2), or (*b*) we choose the natural causal structure defined on U by its structure as a submanifold (with the restricted metric) of X; similarly for V. With choice (*a*) the proposition (0·2·2) is seen to be false by taking x in the set T in Example 6·2. It is still, however, false with choice (*b*). Indeed the observation that the null-cones on X determine its conformal structure would lead us to expect that, unless X is conformally flat, its causal structure will differ, even locally, from that of L^n. To illustrate this difference, let us agree to call a set of nine distinct points in a causal space a *complete square* if they can be labelled a_{mn} $(m, n = 1, 2, 3)$ in such a way as to satisfy the 18 conditions

$$a_{ij} \to a_{i\,j+1}, \quad a_{ij} \to a_{i\,j+2},$$

$$a_{ij} \to a_{i+1\,j}, \quad a_{ij} \to a_{i+2\,j};$$

and let us agree to call a set of eight distinct points a *defective square* if, by suitable labelling, they satisfy the 14 of these conditions which do not involve a_{22}. If V is a non-void open set in L^n, we can always find a non-void open subset V' of V such that every defective square in V' can be completed by the addition of a point in V'; but, in general, if x is a point of a conformally curved manifold, no sufficiently small open neighbourhood U of x contains any non-void open subset U' with the corresponding property (with respect to choice (*b*) of causal structure). This contradicts (0·2·2).

0·3. *Introduction (continued).* In each causal space any two of the causal relations trivially determine the third; but all three relations can be constructed from just one of their number if suitable additional conditions are imposed on the space: the precise condition depending, of course, on the chosen construction. Three constructions which appear particularly natural in the context of an abstract causal space are discussed in section 2. (Those described in 2·1 and 2·3 were noted earlier by Zeeman (13) for L^n.) Starting respectively from the horismos, chronology, and causality, these 'natural' constructions lead to three classes of causal spaces which we call the $\mathfrak{A}$-, $\mathfrak{B}$-, and $\mathfrak{C}$-*spaces*. They are not necessarily the most natural in every context: in particular a manifold equipped with its natural causal structure may fail to be an $\mathfrak{A}$-space or a $\mathfrak{B}$-space; but the cost of such failure is the loss of some 'physically reasonable' properties. A plane-wave solution of the Einstein equations (10) exemplifies a manifold which is not a $\mathfrak{B}$-space in its natural structure, though a $\mathfrak{B}$-space can be constructed from its chronology (Example 6–7).

Section 3 introduces *beams*. A (proper) beam may be approximately described as the union of a family of overlapping subsets of a causal space, each of which is linearly ordered by the horismos: in a space-time manifold it corresponds to a null geodesic of maximal extent. (The complete-square condition referred to in 0·2 is essentially a statement about null geodesics.)

In a sufficiently restricted causal space we may expect the set $C^+(a)$ of all points causally but not chrononologically preceded by a given point a to share some of the 'bounding' properties of the future null-cone of a point in L^n. Section 4—which is concerned with *full* causal spaces (a class of space, defined in 1·3, which includes manifolds in their natural causal structure)—establishes conditions for $C^+(a)$ to be the boundary, with respect to a natural topology, of the set of all points chronologically preceded by a, together with corresponding results for the set $E^+(A)$: defined analogously to $C^+(A)$ when the point a is replaced by a subset A of the space.

In section 5 these bounding properties are considered from a different point of view: we discuss conditions under which a (suitably extensive) set of points linearly ordered by the causality necessarily intersects any set of the type $C^+(a)$ or $E^+(A)$.

Finally, ten illustrative examples are collected together in section 6.

1. *Definitions.* 1·1. We use ' $\supset$ ' (or 'contains', 'is larger than') and related expressions reflexively: i.e. $X \supset X$ for every X. '$X \sim Y$' denotes the set $\{x: x \in X \text{ and } x \notin Y\}$. A set is *non-trivial* if it contains more than one point. A *relation* R on a set X is a subset of $X \times X$, and we write 'xRy' for '$(x, y) \in R$'. Consequently, if R, S are relations, '*R is larger than S*' means that xRy whenever xSy. If $X \supset Y$ and R is a relation on X, the *restriction* of R to Y is the relation $R \cap Y \times Y$ on Y and will be written 'R_Y'.

A relation R on a specified set X will be called *reflexive* if xRx for each x in X, *anti-reflexive* if xRx for no x. We shall call R a *partial ordering* if (i) xRz whenever xRy and yRz and (ii) $x = y$ whenever xRy and yRx; '*R linearly orders Y*' means that R_Y is a partial ordering such that either xRy or yRx whenever x, y are distinct members of Y and '*R orders x, y*' that xRy or yRx or $x = y$. The reflexive relation R will be called *horismotic* if, whenever $(x_i)_{1 \leqslant i \leqslant n}$ is a finite sequence such that $x_i R x_{i+1}$ for each

i (except n) and h, k are integers satisfying $1 \leqslant h \leqslant k \leqslant n$, then (i) $x_1 R x_n$ implies $x_h R x_k$, and (ii) $x_n R x_1$ implies $x_h = x_k$. Any reflexive partial ordering is horismotic.

1·2. The *quadruple* $(X, \prec, \ll, \to)$ will be called a *causal space* if X is a set and $\prec$, $\ll, \to$ are three relations on X satisfying, for each x, y, z in X, the following conditions:

(I) $x \prec x$;

(II) *if* $x \prec y$ *and* $y \prec z$, *then* $x \prec z$;

(III) *if* $x \prec y$ *and* $y \prec x$, *then* $x = y$;

(IV) *not* $x \ll x$;

(V) *if* $x \ll y$ *then* $x \prec y$;

(VI$^+$) *if* $x \prec y$ *and* $y \ll z$, *then* $x \ll z$;

(VI$^-$) *if* $x \ll y$ *and* $y \prec z$, *then* $x \ll z$;

(VII) $x \to y$ *if and only if* $x \prec y$ *and not* $x \ll y$.

(For convenience we shall in future write '$x \prec y \ll z$' for '$x \prec y$ and $y \ll z$', etc.) The set X will be called the *underlying set*; and the three *causal relations* $\prec \ll, \to$ we respectively call the *causality*, *chronology*, and *horismos*. We denote $(X, \prec, \ll, \to)$ by 'X' when no confusion is likely.

An immediate consequence of the axioms is that, whereas the causality is a reflexive and the chronology an anti-reflexive partial ordering, the horismos is a horismotic relation. We mention another consequence for future reference:

LEMMA 1–1. *Let* x, y, z *be points in a causal space. If* $x \prec y \prec z$ *and* $x \to z$, *then* $x \to y \to z$.

Equivalent axiom-systems for the causal relations can be formulated by specifying the properties of any two of them (as in I–VI) and defining the third in terms of the chosen pair (as in VII).

If Y is a subset of X and $(X, \prec, \ll, \to)$ is a causal space, then so is $(Y, \prec_Y, \ll_Y, \to_Y)$, and we shall call this the *causal subspace* of X on Y, or the *restriction* of the space X to the set Y.

Expressions such as 'x is a lower bound for A' or 'y is the first point of B', implying an *unspecified* ordering of a causal space, will always refer to the causality, not the chronology. In particular, 'x (*strictly*) *precedes* y' means '$x \prec y$ (and $x \neq y$)' and 'G is a *chain*' means 'G is a set linearly ordered by the causality'. A set linearly ordered by the chronology will be called a *chronological chain*. The sets $\{x: p \prec x \prec q\}$ and $\{x: s \ll x \ll t\}$ will be denoted respectively by '$[p, q]$' and '$\langle s, t \rangle$'; we shall write '$x \| y$' if neither $x \prec y$ nor $y \prec x$; and a set of which no two distinct points are ordered by the causality will be called *spacelike*.

Each point a of a causal space has a *causal future*, a *chronological future*, and a *null future*, defined respectively as the sets

$$J^+(a) = \{x: a \prec x\},$$
$$I^+(a) = \{x: a \ll x\},$$
$$C^+(a) = \{x: a \to x\} = J^+(a) \sim I^+(a),$$

and a *causal*, a *chronological*, and a *null past*, defined respectively as

$$J^-(a) = \{x\colon x \prec a\},$$
$$I^-(a) = \{x\colon x \ll a\},$$
$$C^-(a) = \{x\colon x \to a\} = J^-(a) \sim I^-(a).$$

Evidently every causal space and every definition or theorem concerning a causal space has a *dual*, obtained by replacing each of three causal relations, R, by its inverse, $R^{-1} = \{(x, y)\colon (y, x) \in R\}$; and we shall indicate this, as we did above, by an interchange of the words 'future' and 'past' and of the superscripts $^+$ and $^-$. Dual definitions and results will normally be treated as self-evident.

If A is a subset of a causal space, the *causal*, *chronological*, and *null futures* of A are defined to be, respectively,

$$J^+[A] = \cup\{J^+(a)\colon a \in A\},$$
$$I^+[A] = \cup\{I^+(a)\colon a \in A\},$$
$$E^+(A) = J^+[A] \sim I^+[A].$$

A causal space will be called *regular* if it satisfies the following condition:

If x_1, x_2, y_1, y_2 are distinct points and $x_i \to y_j$ for each i and j, then the horismos orders x_1, x_2 if and only if it orders y_1, y_2.

Although this is a condition on the horismos alone, Lemma 1–1 enables it to be conveniently re-stated as follows:

If x_1, x_2, y_1, y_2 are distinct points and $x_i \to y_j$ for each i and j, then $x_1 \| x_2$ if and only if $y_1 \| y_2$.

Each subspace of a regular space is regular.

It will be convenient to make use of the definitions given above for the symbols '$[p, q]$', '$\langle s, t\rangle$', '$x \| y$', and '$I^{\pm}(a)$', as well as the concept of duality, whenever we are dealing with a set —not necessarily the underlying set of a given causal space—on which a relation denoted by the symbol '$\prec$' or '$\ll$' has been specified.

1·3. Let $\to$, $\prec$ be two relations on a specified set X. We shall say that $\to$ is a *horismos compatible* (resp. *regularly compatible*) *with the causality* $\prec$ if, for some relation $\ll$, the quadruple $(X, \prec, \ll, \to)$ is a (regular) causal space. The set of all horismoi (regularly) compatible with the causality $\prec$, for the specified X, will be written

$$\{(\text{reg.})\,\text{hor}|\text{cau} \prec\}.$$

The sets $\{\text{hor}|\text{chr} \ll\}$, $\{\text{reg. chr}|\text{cau} \prec\}$, etc., are all similarly defined.

1·4. Let X be a set and $\ll$ an anti-reflexive partial ordering on X. We shall call $\ll$ *future-reflecting* if $I^-(x) \subset I^-(y)$ whenever $I^+(x) \supset I^+(y)$. (Dually, $\ll$ is *past-reflecting* if $I^+(x) \supset I^+(y)$ whenever $I^-(x) \subset I^-(y)$. These are independent properties: see Example 6–5.) We call $\ll$ *weakly distinguishing* if $x = y$ whenever both $I^+(x) = I^+(y)$ and $I^-(x) = I^-(y)$, *future-distinguishing* if $x = y$ whenever $I^+(x) = I^+(y)$. We call $\ll$ *full* if it satisfies both the following condition and its dual:

For each x in X there exists p such that $p \ll x$; and, if $p_1 \ll x$ and $p_2 \ll x$, there exists q such that $p_1 \ll q, p_2 \ll q$, and $q \ll x$.

We define the *Alexandrov topology* $\mathscr{T}^*$ on a set X equipped with an anti-reflexive partial ordering $\ll$ as the smallest (coarsest) topology on X in which each set $I^+(x)$ and

each set $I^-(x)$ is open. If $Y \subset X$, we denote the closure, interior, and boundary of Y with respect to this topology by $Y^{\mathrm{cl}*}$, $Y^{\mathrm{int}*}$, and $Y^{\mathrm{bdy}*}$, respectively. The ordering $\ll$ is full if and only if the collection $\{\langle x, y\rangle : x \ll y\}$ does not contain the empty set and is a base for $\mathscr{T}^*$. (Further characterizations of fullness in terms of $\mathscr{T}^*$ are contained in Lemma 4–1.) There are a number of easily proved relations between the separation-axioms satisfied by the topological space $(X, \mathscr{T}^*)$ and the properties of $\ll$ defined in the preceding paragraph. We mention one, of which we can make use:

LEMMA 1–2. *If $(X, \mathscr{T}^*)$ is Hausdorff and $\ll$ is full, then $\ll$ is future- and past-distinguishing.*

Observe also that, since the ordering $\ll$ is anti-reflexive, no finite subcollection of $\{\langle x, y\rangle : x \ll y\}$ can cover X; consequently, if $\ll$ is full, the topological space $(X, \mathscr{T}^*)$ is not compact.

1·5. When discussing a causal space, we use '$\mathscr{T}^*$' to denote the Alexandrov topology which the chronology defines on the underlying set, and we shall label the causal space 'weakly distinguishing', 'full', etc., if its chronology has the corresponding property.

We shall also need to specify a topology on the underlying set of a causal space in terms of its causality. This will be written '$\mathscr{T}^+$' and is defined as the smallest topology in which each set $J^-(x)$ is closed.

[If X is a manifold equipped with its natural causal structure, X is full and the topology $\mathscr{T}^*$ is smaller than the manifold topology $\mathscr{T}^{\mathrm{man}}$. Lemma 1–2 shows that, if X is not both future- and past-distinguishing, $\mathscr{T}^*$ is strictly smaller (compare Example 6–3); but the converse is false. (Consider the manifold obtained—in the notation of section 6—from the subspace

$$\{x: x_0 \leqslant 1 \quad \textit{and} \quad x_1 > -1\} \cup \{x: x_0 \geqslant -1 \quad \textit{and} \quad x_1 < 1\} \sim \{(1, 1), (-1, -1)\}$$

of $\mathbf{R}^2$ by identifying the two points $(1, x_1)$ and $(-1, -x_1)$ whenever $x_1 > 1$ and defining the null-cones by the metric form $dx_0^2 - dx_1^2$. For this example we are indebted to S. W. Hawking.) However, the following statements are equivalent: (1) $\mathscr{T}^* = \mathscr{T}^{\mathrm{man}}$; (2) *the topological space $(X, \mathscr{T}^*)$ is Hausdorff*; (3) *if each point of $I^-(a)$ precedes each point of $I^+(b)$*, then *b cannot strictly precede a*; (4) *if $u \to v$ and each point of $I^-(v)$ chronologically precedes each point of $I^+(u)$, then $u = v$*; (5) *each point of X possesses an arbitrarily small open neighbourhood—with respect to $\mathscr{T}^{\mathrm{man}}$—whose natural causal structure as a submanifold of X coincides with its structure as a causal subspace of X*. Example 6–2 shows that $\mathscr{T}^+$ need not be comparable with $\mathscr{T}^{\mathrm{man}}$ or $\mathscr{T}^*$: we revert to this in Theorem 4–5. Since $\mathscr{T}^* \subset \mathscr{T}^{\mathrm{man}}$, the space $(X, \mathscr{T}^{\mathrm{man}})$ cannot be compact.]

2. *Construction from a single relation.* 2·1. *Construction from the horismos.* Let X be a set and $\to$ a horismotic relation on X. We define two further relations $\prec^{\mathfrak{A}}$, $\ll^{\mathfrak{A}}$ on X in terms of $\to$ as follows:

$$\left.\begin{array}{l} (a)\ x \prec^{\mathfrak{A}} y \textit{ if and only if there exists a finite sequence } (u_i)_{1 \leqslant i \leqslant n} \textit{ satisfying} \\ \qquad\qquad x = u_1 \to u_2 \to \ldots \to u_n = y; \\ (b)\ x \ll^{\mathfrak{A}} y \textit{ if and only if } x \prec^{\mathfrak{A}} y \textit{ and not } x \to y. \end{array}\right\} \quad (2{\cdot}1{\cdot}1)$$

Then $(X, \prec^{\mathfrak{A}}, \ll^{\mathfrak{A}}, \to)$ is easily seen to be a causal space; so

$$\prec^{\mathfrak{A}} \in \{\text{cau}|\text{hor}\to\}, \quad \ll^{\mathfrak{A}} \in \{\text{chr}|\text{hor}\to\}.$$

Since, for any $\prec$ in $\{\text{cau}|\text{hor}\to\}$, $x \prec y$ whenever $x \prec^{\mathfrak{A}} y$, it follows that

$$\prec^{\mathfrak{A}} = \cap\{\text{cau}|\text{hor}\to\}, \quad \ll^{\mathfrak{A}} = \cap\{\text{chr}|\text{hor}\to\}. \tag{2·1·2}$$

Now let $(X, \prec, \ll, \to)$ be any causal space: so that $\to$ is certainly a horismotic relation. We shall call this an $\mathfrak{A}$-*space* if $\prec$, $\ll$ coincide with the relations $\prec^{\mathfrak{A}}$, $\ll^{\mathfrak{A}}$ defined in (2·1·1). By (2·1·2) a condition sufficient for the causal space $(X, \prec, \ll, \to)$ to be an $\mathfrak{A}$-space is that $x \prec^{\mathfrak{A}} y$ whenever $x \prec y$. Example 6–1 illustrates that, even in a very simple $\mathfrak{A}$-space, the number n in (2·1·1) must be allowed to take arbitrarily large values.

[[Each sufficiently small open neighbourhood U of an arbitrary point in a manifold X is an $\mathfrak{A}$-space with respect to the causal relations naturally determined on U by its submanifold structure in X. The remarks in 1·5, together with a simple compactness argument, therefore establish that X in its natural causal structure is an $\mathfrak{A}$-space if $(X, \mathscr{T}^*)$ is Hausdorff.]]

2·2. *Construction from the chronology.* Let X be a set and $\ll$ an anti-reflexive partial ordering on X. We define two further relations $\prec^{\mathfrak{B}}$, $\to^{\mathfrak{B}}$ on X in terms of $\ll$ as follows:

$$\left.\begin{array}{l} (a)\ x \prec^{\mathfrak{B}} y \text{ if and only if } I^+(x) \supset I^+(y) \text{ and } I^-(x) \subset I^-(y); \\ (b)\ x \to^{\mathfrak{B}} y \text{ if and only if } x \prec^{\mathfrak{B}} y \text{ and not } x \ll y. \end{array}\right\} \tag{2·2·1}$$

If $\prec \in \{\text{cau}|\text{chr}\ll\}$, then $x \prec^{\mathfrak{B}} y$ whenever $x \prec y$. On the other hand if $x \prec^{\mathfrak{B}} y$, we can find $\prec$ in $\{\text{cau}|\text{chr}\ll\}$ such that $x \prec y$. (Define $\prec$ thus: $u \prec v$ if and only if either $u = v$ or $u \ll v$ or $u = x$, $v = y$.) This establishes that

$$\prec^{\mathfrak{B}} = \cup\{\text{cau}|\text{chr}\ll\}, \quad \to^{\mathfrak{B}} = \cup\{\text{hor}|\text{chr}\ll\}. \tag{2·2·2}$$

The quadruple $(X, \prec^{\mathfrak{B}}, \ll, \to^{\mathfrak{B}})$ does not, however, necessarily satisfy axiom III for a causal space: we easily verify that it is a causal space—i.e. that

$$\prec^{\mathfrak{B}} \in \{\text{cau}|\text{chr}\ll\}, \quad \to^{\mathfrak{B}} \in \{\text{hor}|\text{chr}\ll\}$$

—if and only if the ordering $\ll$ is weakly distinguishing.

Now let $(X, \prec, \ll, \to)$ be any causal space: so that $\ll$ is certainly an anti-reflexive partial ordering. We shall call this a $\mathfrak{B}$-*space* if $\prec$, $\to$ coincide with the relations $\prec^{\mathfrak{B}}$, $\to^{\mathfrak{B}}$ defined in (2·2·1). Evidently a $\mathfrak{B}$-space is weakly distinguishing; and by (2·2·2) a condition sufficient for the causal space $(X, \prec, \ll, \to)$ to be a $\mathfrak{B}$-space is that $x \prec y$ whenever $x \prec^{\mathfrak{B}} y$.

2·3. *Future-reflecting* $\mathfrak{B}$-*spaces.* Suppose X is a set, $\ll$ an anti-reflexive ordering on X, and $\prec^{\mathfrak{B}+}$ the relation on X defined by

$$x \prec^{\mathfrak{B}+} y \quad \textit{if and only if} \quad I^+(x) \supset I^+(y).$$

The relation $\prec^{\mathfrak{B}}$, defined by (2·2·1), coincides with $\prec^{\mathfrak{B}+}$ if and only if $\ll$ is future-reflecting; and, since $\prec^{\mathfrak{B}+}$ is larger than $\prec^{\mathfrak{B}}$, the relation $\prec^{\mathfrak{B}+}$ cannot be a causality

compatible with the chronology $\ll$ if it does not coincide with $\prec^{\mathfrak{B}}$. It follows that a necessary and sufficient condition for $\prec^{\mathfrak{B}+}$ to be a causality compatible with the chronology $\ll$ is that $\ll$ should be both future-reflecting and weakly distinguishing. Correspondingly the causality of a given causal space will coincide with the relation $\prec^{\mathfrak{B}+}$ defined from its chronology if and only if it is a future-reflecting $\mathfrak{B}$-space: a sufficient condition for this is that $x \prec y$ whenever $x \prec^{\mathfrak{B}+} y$.

(An anti-reflexive partial ordering is future-reflecting and weakly distinguishing. if and only if it is future-reflecting and future-distinguishing. If we postulate only that $\ll$ is future-distinguishing, the relation $\prec^{\mathfrak{B}+}$ satisfies every condition for a causality compatible with the chronology $\ll$ except axiom VI$^-$. This might suggest weakening the axioms for a causal space by omitting one of the dual axioms VI$^+$ and VI$^-$: indeed, to be either future- or past-distinguishing seems a most natural requirement for any reasonable event-space, whereas some fairly reasonable spaces are neither future- nor past-reflecting—see Examples 6–4 and 6–6. However, an axiom system reduced in this way appears to be insufficient for results analogous to many of those considered in this paper—in particular those of section 3.)

2·4. *Construction from the causality.* Let X be a set and $\prec$ a reflexive partial ordering on X. We define two further relations $\rightarrow^{\mathfrak{C}}$, $\ll^{\mathfrak{C}}$ on X in terms of $\prec$ as follows:

$$\left.\begin{array}{ll} (a) & x \rightarrow^{\mathfrak{C}} y \textit{ if and only if } \text{(i) } x \prec y \textit{ and (ii) } \prec \textit{linearly orders } [u,v] \textit{ whenever} \\ & [u,v] \textit{ is a proper subset of } [x,y]; \\ (b) & x \ll^{\mathfrak{C}} y \textit{ if and only if } x \prec y \textit{ and not } x \rightarrow^{\mathfrak{C}} y. \end{array}\right\} \quad (2\cdot4\cdot1)$$

Then $(X, \prec, \ll^{\mathfrak{C}}, \rightarrow^{\mathfrak{C}})$ is a causal space; so

$$\rightarrow^{\mathfrak{C}} \in \{\text{hor}|\text{cau} \prec\}, \quad \ll^{\mathfrak{C}} \in \{\text{chr}|\text{cau} \prec\};$$

but since $\prec$ is itself a horismos and $\varnothing$ a chronology compatible with the causality $\prec$, it follows that the new relations cannot be expressed precisely as in (2·1·2) and (2·2·2). Instead we find:

$$\rightarrow^{\mathfrak{C}} = \cup\{\text{reg. hor}|\text{cau} \prec\}, \quad \ll^{\mathfrak{C}} = \cap\{\text{reg. chr}|\text{cau} \prec\}. \quad (2\cdot4\cdot2)$$

To prove this we note first that, if $\rightarrow$ belongs to $\{\text{reg. hor}|\text{cau} \prec\}$ and $x \rightarrow y$, then $x \rightarrow^{\mathfrak{C}} y$: for if in this case $[u,v]$ is a subset of $[x,y]$ and is not linearly ordered by $\prec$, there exist p, q such that $p \| q$, $x \rightarrow p$, $u \rightarrow p$, $x \rightarrow q$, $u \rightarrow q$, which, by the assumed regularity, implies that $x = u$; and similarly $y = v$. On the other hand, if $x \rightarrow^{\mathfrak{C}} y$, we can find $\rightarrow$ in $\{\text{reg. hor}|\text{cau} \prec\}$ satisfying $x \rightarrow y$. (Define $\rightarrow$ thus: $u \rightarrow v$ if and only if either $u = v$ or $x \prec u \prec v \prec y$.) This proves the first part of (2·4·2); the second follows trivially.

Now let $(X, \prec, \ll, \rightarrow)$ be any causal space: so that $\prec$ is certainly a reflexive partial ordering. We shall call this a $\mathfrak{C}$-*space* if $\rightarrow$, $\ll$ coincide with the relations $\rightarrow^{\mathfrak{C}}$, $\ll^{\mathfrak{C}}$ defined in (2·4·1). A $\mathfrak{C}$-space need not be regular—Example 6–9—but by (2·4·2) a condition sufficient for the regular causal space $(X, \prec, \ll, \rightarrow)$ to be a $\mathfrak{C}$-space is that $x \rightarrow y$ whenever $x \rightarrow^{\mathfrak{C}} y$, or equivalently that $x \ll^{\mathfrak{C}} y$ whenever $x \ll y$. That this condition is not sufficient when X is not regular is illustrated by Example 6–6 or by again

observing that $(X, \prec, \varnothing, \prec)$ is a causal space when $\prec$ is any reflexive partial ordering on X.

[[If X is a manifold equipped with its natural causal structure, then $x \ll^{\mathfrak{C}} y$ whenever $x \ll y$. Since X is regular (see 3·2), X is a $\mathfrak{C}$-space.]]

3. *Girders, beams, and links.* 3·1. *Girders and hypergirders.* Let G be a finite chain of N points in the causal space X, and arrange the points of G in their ordering by $\prec$: i.e. construct the finite sequence $(g_i)_{1 \leqslant i \leqslant N}$ for which each g_i belongs to G and (if $i < N$) strictly precedes g_{i+1}. We shall call G a *girder* if (i) $N \geqslant 3$ and (ii) $g_i \to g_j$ whenever $j - i = 2$. By Lemma 1–1 this implies that $g_i \to g_j$ whenever $j - i = 1$.

$$\begin{array}{ccccccccc} \cdots & g_{i-2} & \to & g_i & \to & g_{i+2} & \cdots \\ & \searrow & \nearrow & \searrow & \nearrow & & \\ \cdots\cdots & & g_{i-1} & \to & g_{i+1} & & \cdots\cdots \end{array}$$

It will frequently be convenient to denote the girder G by the symbol for the N-tuple $(g_1, \ldots, g_N)$ to exhibit its structure. For example, the following three statements are equivalent: (1) (a, b, c) *is a girder*; (2) $\{a, b, c\}$ *is a girder and* $a \prec b \prec c$; (3) a, b, c *are distinct and* $a \to b \to c$ *and* $a \to c$.

The following lemma is an immediate consequence of the definition and Lemma 1–1:

LEMMA 3–1. *If* $(g_1, \ldots, g_N)$ *is a girder and* h *lies strictly between* g_r *and* g_{r+1}, *then* $(g_1, \ldots, g_r, h, g_{r+1}, \ldots, g_N)$ *is a girder.*

A subset H of X will be called a *hypergirder* if (i) $H \neq \varnothing$, and (ii) for any x, y in H there exists a girder G such that $x, y \in G$ and $G \subset H$. In particular, every girder is a hypergirder; also any hypergirder is a chain of at least three points.

The defining property of a hypergirder generalizes to the following result:

LEMMA 3–2. *If* F *is any finite subset of the hypergirder* H, *there exists a girder* G *such that* $F \subset G \subset H$.

Proof. We can suppose the lemma proved when the subset has n points, for some $n \geqslant 2$. Let F be a set of $n+1$ points in H; choose f in $F \sim \{\min F, \max F\}$; choose a girder G' such that $F \sim \{f\} \subset G' \subset H$; and let $G = G' \cup \{f\}$. Then either G equals G', or G is a girder by Lemma 3–1.

3·2. *Proper beams.* A subset of a causal space X will be called a *proper beam* if it is a maximal hypergirder—i.e. a hypergirder which is not a proper subset of any hypergirder in X. Zorn's lemma then yields:

THEOREM 3–3. *Every hypergirder is contained in some proper beam.*

The points x and y will be called *proximate* if they are successive points of some girder —i.e. if there exists a girder (g_1, g_2, g_3) such that $x = g_i$, $y = g_j$, where $|i - j| = 1$. We can now state the principal theorem of this section.

THEOREM 3–4. *Each pair of proximate points belong to precisely one proper beam if and only if the causal space is regular.*

To prove this we first establish a lemma.

LEMMA 3–5. *Let* $G = (g_1, \ldots, g_N)$ *be a girder and* B *a proper beam which contains* $\{g_i, g_{i+1}\}$ *for some* i. *Then, if the space is regular,* B *contains* G.

Proof of lemma. Let the space be regular and suppose g_{i+2} is defined (i.e. $i < N-1$) but does not belong to B. Since B is maximal, $B' = B \cup \{g_{i+2}\}$ is not a hypergirder; so there exists p in B such that no girder containing $\{p, g_{i+2}\}$ is contained in B'. By Lemma 3–2, B contains a girder $H = (h_1, \ldots, h_M)$ containing $\{p, g_i, g_{i+1}\}$; by hypothesis $g_{i+2} \notin H$. Let m be the largest integer such that $h_m \to g_{i+2}$: by Lemma 1–1 it then follows that $h_{m-1} \to g_{i+2}$. If $m = M$, then $(h_1, \ldots, h_M, g_{i+2})$ is a girder contained in B': a contradiction. If $m \neq M$, then h_{m-1}, h_m, h_{m+1}, g_{i+2} are distinct points for which $h_{m-1} \to h_{m+1}, h_m \to h_{m+1}, h_{m-1} \to g_{i+2}, h_m \to g_{i+2}$: since the space is regular and $h_{m-1} \to h_m$ but not $h_{m+1} \to g_{i+2}$, it follows that $g_{i+2} \to h_{m+1}$; consequently (by Lemma 3–1) $(h_1, \ldots, h_m, g_{i+2}, h_{m+1} \ldots, h_M)$ is a girder contained in B': the same contradiction. Similarly g_{i-1}, if defined, must belong to B, and the lemma follows by finite induction.

Proof of Theorem 3–4. Suppose the space is regular and that p, q are successive points of the girder $G = (g_1, g_2, g_3)$. By Theorem 3–3 there exists a proper beam B containing G. Let B' be any proper beam containing $\{p, q;\}$ then $B' \supset G$ by Lemma 3–5. Let $x \in B$. By Lemma 3–2 there exists a girder $H = (h_1, \ldots, h_M)$ such that $G \cup \{x\} \subset H \subset B$. Let $g_1 = h_i$, $g_2 = h_j$, $g_3 = h_k$, where $i < j < k$. If $i+1 = j$, then $h_{i+1} \in B'$. Suppose $i+1 < j$. Since (g_1, g_2, g_3) is a girder and H a chain, Lemma 3–1 implies that

$$K = (h_i, h_{i+1}, h_j, h_k)$$

is a girder; and since h_j and h_k belong to B', we deduce from Lemma 3–5 that $B' \supset K$— so in this case, too, $h_{i+1} \in B'$. Since both h_i and h_{i+1} belong to B', Lemma 3–5 implies that $B' \supset H$; hence $x \in B'$. Therefore $B \subset B'$, so that, since B is maximal, $B' = B$.

Suppose conversely that each pair of proximate points belong to exactly one proper beam. Let x_1, x_2, y_1, y_2 be distinct points such that $x_i \to y_j$ for each i, j, and suppose $y_1 \prec y_2$. Then y_1 and y_2 are proximate; and if B is the unique proper beam containing $\{y_1, y_2\}$, B must, by Theorem 3–3, contain both the girders (x_1, y_1, y_2) and (x_2, y_1, y_2): therefore not $x_1 \| x_2$. Similarly if $x_1 \prec x_2$, then not $y_1 \| y_2$. It follows that the space is regular, which concludes the proof.

Example 6–1 illustrates why 'proximate' cannot be weakened to 'ordered by the horismos' in Theorem 3–4.

[[Let X be a manifold with its natural causal structure. Each proper beam then coincides with a (not necessarily maximal) null geodesic. (It is convenient here to identify a null geodesic with the point-set which it determines in the manifold.) We may deduce that X is necessarily regular. Conversely, if G is a null geodesic and $x \in G$, there exists an open neighbourhood U of x such that $U \cap G$ is linearly ordered by the horismos belonging to the natural causal structure on U as a submanifold of X. By the remarks in 1·5 and a compactness argument it follows that, if $(X, \mathscr{T}^*)$ is Hausdorff, G is a hypergirder. We conclude, using Theorem 3–3, that if $(X, \mathscr{T}^*)$ is Hausdorff, the proper beams and maximal null geodesics coincide. Example 6–3 shows that if $(X, \mathscr{T}^*)$ is not Hausdorff, this result can fail.]]

3·3. *Beams.* Proper beams constitute a sometimes inconveniently restricted class (see, for example, 3·4). We introduce a simple generalization, which will coincide with the original concept in all but the most pathological spaces. [[The definitions coincide, in particular, for any manifold equipped with its natural causal structure.]]

We shall call a set B in a causal space a *beam* if either of the following conditions is satisfied:

(1) *B is a proper beam;*

(2) *$B = \{a, b\}$, where a, b are distinct points ordered by the horismos, and B is not contained in any girder.*

(Condition (2) implies both that a and b are not proximate and that $[a, b] \sim \{a, b\}$ is empty.)

Let P be any non-trivial set linearly ordered by the horismos, and define

$$K = \cup \{[p, q] : p, q \in P\}.$$

In the set K (partially ordered by $\prec$) choose a maximal chain H containing the chain P. If H has more than two points, it is a hypergirder; using Theorem 3–3 we conclude:

THEOREM 3–6. *In any causal space each non-trivial set linearly ordered by the horismos is a subset of some beam.*

3·4. *Generators of a null future.* Any regular causal space has a geometric property characteristic of Minkowski space-time, conveniently illustrated by taking A (in the following paragraph) to be the subset

$$\{(0, x_1, x_2) : 1 \leqslant x_1^2 + x_2^2 \leqslant 2 \quad \text{or} \quad x_1 = x_2 = 0\}$$

of the causal space $\mathbf{L}^3$.

Let A be a spacelike subset of a regular causal space, so that $E^+(A)$ contains A. Define

$$A_0 = \{x : x \in A \text{ and } x \text{ strictly precedes some point of } E^+(A)\},$$

$$S = E^+(A) \sim (A \sim A_0);$$

and call $B \cap S$ a *generator* of S whenever B is a beam intersecting S more than once. Then the following is a straightforward consequence of Theorems 3–4 and 3–6: S is the union of its generators; each generator has a first point in A_0 and may or may not have a last point, the set of all last points being spacelike; and any point common to two distinct generators is either the first or the last point of both.

3·5. *Links and linked chains.* The subset $L = [p, q]$ of a causal space will be called a *link* if L is a non-trivial chain and $p \to q$. By 1–1 and 3–6 every link is a subset of some beam.

Let Y be a subset of a causal space. We shall call the finite sequence $(x_i)_{1 \leqslant i \leqslant n}$ (where $n \geqslant 1$) a *linkage of Y from x_1 to x_n* if, for each i (except n), the set $[x_i, x_{i+1}]$ is a link contained in Y. We call Y *linked* if a linakge of Y from p to q exists whenever p precedes q in Y. If p strictly precedes q in a linked chain R, then

$$R \cap [p, q] = \cup \{[x_i, x_{i+1}] : 1 \leqslant i < n\},$$

where $(x_i)_{1 \leqslant i \leqslant n}$ is a linkage of R from p to q.

In a regular space, if $(g_1, \ldots, g_N)$ is a girder, then each set $[g_i, g_{i+1}]$ is a link ; consequently any beam is a linked chain. Therefore, by Theorem 3–6:

LEMMA 3–7. *In a regular space any two points ordered by the horismos belong to some linked chain.*

COROLLARY 3–8. *A regular space is an $\mathfrak{A}$-space if and only if its underlying set is linked.*

3·6. *Note.* Girders, hypergirders, and beams are all concepts which could evidently be defined for a set X equipped only with a horismotic relation $\rightarrow$. Where we have in fact used a partial ordering on X, we would then employ the ordering $\prec^{\mathfrak{A}}$ defined in (2·1·1).

4. *Full causal spaces.* 4·1. The first lemma characterizes full orderings in terms of $\mathscr{T}^*$-closure.

LEMMA 4–1. *Let $\ll$ be an anti-reflexive partial ordering on a set X. Then the following are equivalent statements*:

(1) *$\ll$ is full*;

(2) *$x \in I^+(x)^{\mathrm{cl}^*} \cap I^-(x)^{\mathrm{cl}^*}$ for each x in X*;

(3) *$I^+(x)^{\mathrm{cl}^*} = \{y : I^+(x) \supset I^+(y)\}$ and*

$$I^-(x)^{\mathrm{cl}^*} = \{y : I^-(x) \supset I^-(y)\}, \quad \text{for each } x \text{ in } X.$$

Proof. (3) trivially implies (2); and, to see that (2) implies (1), observe that, if

$$x \in I^-(x)^{\mathrm{cl}^*}$$

and P is a finite or void subset of $I^-(x)$, then $\cap\{I^+(p) : p \in P\}$ must intersect $I^-(x)$. Finally, assume (1). Any $\mathscr{T}^*$-neighbourhood of y contains a neighbourhood $\langle u, v\rangle$ of y, and $\langle y, v\rangle$ is not empty; so, if $I^+(x) \supset I^+(y)$, then $y \in I^+(x)^{\mathrm{cl}^*}$. Conversely, if $y \in I^+(x)^{\mathrm{cl}^*}$ and $y \ll z$, then $I^-(z)$ is a $\mathscr{T}^*$-neighbourhood of y and so intersects $I^+(x)$: therefore $x \ll z$.

4·2. The next two lemmas contain the results we need concerning full causal spaces.

LEMMA 4–2. *Let A be a subset of a full causal space. Then $J^+[A] \subset I^+[A]^{\mathrm{cl}^*}$ and $I^+[A] = J^+[A]^{\mathrm{int}^*}$.*

Proof. By Lemma 4–1, $J^+(a) \subset I^+(a)^{\mathrm{cl}^*}$ for each a in A, which implies the first result. If $x \in J^+[A]^{\mathrm{int}^*}$, there exist p, q, r such that $x \in \langle p, q\rangle \subset J^+[A]$ and $p \ll r \ll x$; then $x \in I^+[A]$ since $r \in J^+[A]$. As $I^+[A]$ is $\mathscr{T}^*$-open, this concludes the proof.

COROLLARY 4–3. *If A is a given subset of a full causal space, then all sets H satisfying*

$$I^+[A] \subset H \subset J^+[A]$$

have the same closure and the same interior with respect to $\mathscr{T}^*$; and $H^{\mathrm{bdy}^*} \supset E^+(A)$.

LEMMA 4–4. *Let X be a full causal space. Then the following are equivalent statements*:

(1) *X is a future-reflecting $\mathfrak{B}$-space*;

(2) *$J^+(x) = I^+(x)^{\mathrm{cl}^*}$ for each x in X*;

(3) $J^+[A] = I^+[A]^{\mathrm{cl}*}$ *for each* A *compact with respect to* $\mathscr{T}^+$;

(4) $J^+[A]^{\mathrm{bdy}*} = E^+(A)$ *for each* A *compact with respect to* $\mathscr{T}^+$.

Proof. By Lemma 4–1, (2) implies (1). (3) trivially implies (2); while (3) and (4) are equivalent by Lemma 4–2 and its corollary. We therefore assume (1), and let $x \in I^+[A]^{\mathrm{cl}*}$. If F is a finite subset of $I^+(x)$, $K = \cap\{J^-(p): p \in F\}$ contains a $\mathscr{T}^*$-neighbourhood of x and so intersects $I^+[A]$: consequently some point of A precedes some point of K, which implies that $A \cap K$ is not void. It follows that the collection $\mathscr{F} = \{A \cap J^-(p): p \in I^+(x)\}$ has the finite-intersection property; so that, if A is $\mathscr{T}^+$-compact, we can choose a in $\cap\mathscr{F}$. For each q in $I^+(x)$ we can choose p in $\langle x, q\rangle$; then $a \prec p$; therefore $a \ll q$. Consequently $I^+(a) \supset I^+(x)$, which, by (1), establishes that $x \in J^+(a) \subset J^+[A]$. (3) now follows from Lemma 4–2.

From the duals of 4–3 and 4–4 we deduce:

THEOREM 4–5. *Let* X *be a full causal space. Then* $\mathscr{T}^+ \subset \mathscr{T}^*$ *if and only if* X *is a past-reflecting* $\mathfrak{B}$*-space.*

It follows that $\mathscr{T}^*$-compactness of a subset of a full past-reflecting $\mathfrak{B}$-space is sufficient to ensure its $\mathscr{T}^+$-compactness.

4·3. *Note.* We return to the problem, considered in 2·2, of defining a causal structure on a set X equipped with an anti-reflexive partial ordering $\ll$. Suppose we define a relation $\prec^{\mathscr{T}}$ on X in terms of $\ll$ as follows:

$$x \prec^{\mathscr{T}} y \quad \textit{if and only if} \quad y \in I^+(x)^{\mathrm{cl}} \quad \textit{and} \quad x \in I^-(y)^{\mathrm{cl}},$$

where $^{\mathrm{cl}}$ denotes closure with respect to a specified topology $\mathscr{T}$ on X. In order that $\prec^{\mathscr{T}}$ should satisfy axiom I for a causality on X, each point x of X must belong to both $I^+(x)^{\mathrm{cl}}$ and $I^-(x)^{\mathrm{cl}}$, which is easily seen to imply that $\prec^{\mathscr{T}}$ contains $\prec^{\mathfrak{B}}$; so that—by (2·2·2)—no fresh method of construction is to be looked for in this direction. Lemma 4–1 establishes the condition under which $\prec^{\mathscr{T}}$ coincides with $\prec^{\mathfrak{B}}$ in the case $\mathscr{T} = \mathscr{T}^*$: namely, $\prec^{\mathscr{T}^*} = \prec^{\mathfrak{B}}$ if and only if $\ll$ is full. Thus the causality of a causal space X coincides with the relation $\prec^{\mathscr{T}^*}$ determined by its chronology if and only if X is a full $\mathfrak{B}$-space.

4·4. *Note.* Let H be a subset of the causal space X. A necessary and sufficient condition for H to satisfy

$$I^+[A] \subset H \subset J^+[A] \quad \textit{for some } A$$

is that $I^+[H] \subset H$. A method which suggests itself of defining a 'boundary' of such a set without resort to a topology on X is to define it as the set

$$B^+(H) = \{x: I^+(x) \subset H \quad \textit{and} \quad I^-(x) \subset X \sim H\}.$$

The assumed condition on H implies that $B^+(H)$ contains $H^{\mathrm{bdy}*}$; if in addition X is full, these two sets are equal. Example 6–9 shows that they need not be equal if X is not full.

5. *Bounding properties of the null future.* 5·1. Let A be a set in a causal space X. To express the condition that a given chain should intersect $E^+(A)$, provided that it is,

roughly speaking, sufficiently extensive to be in a position to do so, we shall define a chain in X to be *future-connecting for* A if it either intersects $E^+(A)$ or does not intersect $J^+[A]$ or does not intersect $X \sim J^+[A]$. Observe that the chain R is future-connecting for A if and only if $R \cap [p,q]$ intersects $E^+(A)$ whenever

$$p \in R \sim J^+[A] \quad \text{and} \quad q \in R \cap J^+[A].$$

If a chain is future-connecting for each one-point subset $\{x\}$ of X, we call it *future-connecting*; and we shall call a causal space *future-connected* if each causally ordered pair of points in it belong to some future-connecting chain.

The causal space X will be called *future-conjunct for* A if $[p,q]$ intersects $E^+(A)$ whenever $p \in X \sim J^+[A]$, $q \in J^+[A]$, and $p \prec q$; the space is *future-conjunct* if it is future-conjunct for each one-point set. (A useful equivalent is: X is future-conjunct if and only if, whenever $x_1 \prec z$, $x_2 \prec z$, $x_1 \| x_2$, there exists y such that $x_1 \to y$, $x_2 \to y$, and $y \prec z$.) Every future-connected space is future-conjunct: Example 6–10 disproves the converse.

5·2. A chain which is future-connecting or space which is future-conjunct is easily seen to have the corresponding property for all finite sets, but not necessarily for arbitrary sets. (Consider the null future of

$$\{x: |x_0| < 1, x_1 = 0\} \quad \text{or} \quad \{x: x_0 = 0, |x_1| < 1\}$$

in L^2.) This suggests introducing a compactness condition; and it is again compactness with respect to $\mathscr{T}^+$ that leads to pleasant results.

Let us call a chain *setwise future-connecting* if it is future-connecting for each set compact with respect to $\mathscr{T}^+$, call a space *setwise future-connected* if each causally ordered pair of points belong to a setwise future-connecting chain, and call a space *setwise future-conjunct* if it is future-conjunct for each set compact with respect to $\mathscr{T}^+$.

Theorem 5–5 will show that each regular future-conjunct space is setwise future-conjunct. On the other hand the future-connecting chain $\{x: |x_0| = x_1 > 0\}$ in L^2 is not future-connecting for the $\mathscr{T}^+$-compact set

$$\{x: 0 = x_0 < x_1 \quad \textit{or} \quad x_0 = x_1 = -1\}. \tag{5·2·1}$$

This particular chain is not maximal; but if we replace L^2 by its restriction to $\{x: x_0 + x_1 \neq 0\}$, the chain

$$\{x: x_0 = x_1 > 0 \quad \textit{or} \quad x_0 < x_1 = 0\} \tag{5·2·2}$$

is future-connecting and maximal, while still not future-connecting for the set (5.2.1). The chain $\{x: x_0 \neq 0 = x_1\}$ in this space illustrates that a maximal chain need not even be future-connecting, although the space is actually setwise future-connected (Theorem 5–4).

The first theorem of this section establishes that for beams and chronological chains the situation is more cheerful. We begin by formulating a condition sufficient to exclude the behaviour exemplified by (5·2·2).

LEMMA 5–1. *Let R be a future-connecting chain such that, whenever S is a non-void subset of R and a is a lower bound for S but not for R, there exists a lower bound h for S such that $a \prec h$ and $h \in R$. Then R is setwise future-connecting.*

Proof. Suppose there exists a $\mathscr{T}^+$-compact set A and points p, q in R such that $p \notin J^+[A]$, $q \in J^+[A]$, and $E^+(A) \cap R = \varnothing$. Let $S = R \cap J^+[A]$, choose a in

$$\cap \{J^-(s) \cap A : s \in S\},$$

and choose a point h with the stated properties. Since $a \prec h \in R$, but $h \notin E^+(A) \cap R$, it follows that $a^* \ll h$ for some a^* in A. Since $p \notin J^+(a^*)$, there exists k in $C^+(a^*) \cap R$, and k must strictly precede h since $a^* \ll h$. Therefore $k \notin S$ (for h bounds S below), which contradicts the definition of S.

THEOREM 5–2. *A future-connecting chain R in a causal space X is setwise future-connecting if either of the following statements is true*:

(1) *R is a chronological chain*;

(2) *X is regular and R is a linked chain.*

Proof. We show that if the future-connecting chain R does not satisfy the condition of Lemma 5–1, then neither (1) nor (2) can hold. Let S be a non-void subset of R; let $J^+(a) \supset S$ and $p \in R \sim J^+(a)$. Suppose S has no lower bounds in $J^+(a) \cap R$, so that, in particular, $a \notin R$. Choose h in $C^+(a) \cap R$: then there exists s in S strictly preceding h. By Lemma 1–1, $s \to h$: therefore R is not a chronological chain. Suppose that $(x_i)_{1 \leqslant i \leqslant n}$ is a linkage of R from p to h, and let $j = \min\{i : a \prec x_i\}$; so $j > 1$. By hypothesis there exists t in S strictly preceding x_j. Since $a \notin R$, t is not a : therefore, by Lemma 1–1, (a, t, x_j) is a girder. The point t cannot precede x_{j-1}; so t and x_j are proximate points in the link $L = [x_{j-1}, x_j]$. Let B be a beam containing L. Supposing that X is regular, $a \in B$ by Theorem 3–4; so, since B is a chain and a neither precedes x_{j-1} nor is strictly preceded by x_j, a must belong to L and therefore to R. This contradiction concludes the proof.

LEMMA 5–3. *Let X be a regular space in which every link is future-connecting, and let R be a linked chain in X. Then R is setwise future-connecting.*

Proof. Let $p \in R \sim J^+(a)$, $q \in R \cap J^+(a)$; let $(x_i)_{1 \leqslant i \leqslant n}$ be a linkage of R from p to q; and let $j = \min\{i : a \prec x_i\}$. Then $C^+(a)$ intersects $[x_{j-1}, x_j]$, which is a subset of R. The result follows from Theorem 5–2.

Obviously every link in a future-conjunct space is future-connecting. Therefore 3–8 and 5–3 yield:

THEOREM 5–4. *If X is a regular $\mathfrak{A}$-space, the following are equivalent statements*:

(1) *X is future-conjunct*;

(2) *each link in X is future-connecting*;

(3) *X is setwise future-connected.*

In addition we can now prove:

THEOREM 5–5. *Every regular future-conjunct space is setwise future-conjunct.*

Proof. Let A be $\mathscr{T}^+$-compact in the regular future-conjunct space X, and let $p \prec q$, $p \notin J^+[A]$, $q \in J^+[A]$. Choose a' in $A \cap J^-(q)$, and let K be a maximal chain in A containing $\{a'\}$. Since A is $\mathscr{T}^+$-compact, K has a first member, a. Then either $a \in [p,q] \cap E^+(A)$ or $a \notin J^+(p)$. Suppose the latter. Since $q \in J^+(p)$, we can find h in $C^+(p) \cap [a,q]$. By Lemma 3–7 there exists a linked chain containing $\{p,h\}$: by Lemma 5–3 this must intersect $E^+(A)$. It follows that $[p,q] \cap E^+(A) \neq \varnothing$.

5·3. We now turn our attention to full spaces. From Lemma 4–4 we have:

THEOREM 5–6. *If R is a chain connected with respect to the topology $\mathscr{T}^*$ in a full future-reflecting $\mathfrak{B}$-space, then R is setwise future-connecting.*

The condition that X be a future-reflecting $\mathfrak{B}$-space can only be weakened at some cost:

THEOREM 5–7. *Let X be a full causal space which is not a future-reflecting $\mathfrak{B}$-space. Then there exists a point in X which is the last point of any future-connecting chronological chain to which it belongs.*

Proof. By Lemmas 4–2 and 4–4 there exist points a and p in X such that $p \in I^+(a)^{\mathrm{cl}^*} \sim J^+(a)$. Let $p \in R$, where R is a future-connecting chronological chain. If there exists q in R such that $p \ll q$, then (since $I^-(q)$ is a $\mathscr{T}^*$-neighbourhood of p) $q \in I^+(a)$: so there exists h in $C^+(a) \cap R$. Then $p \ll h$ (since otherwise $a \prec p$); so we may choose r in $\langle p,h\rangle$. As before it follows that $r \in I^+(a)$; so $a \ll h$: a contradiction.

[[If X is a manifold equipped with its natural causal structure, it follows that, of the conditions below, (1) and (2) are equivalent and imply (3):

(1) *X is a future-reflecting $\mathfrak{B}$-space*;

(2) *every non-space like path in X describes a setwise future-connecting chain*;

(3) *X is setwise future-connected.*

Example 6–2 shows that (3) does not imply (1).]]

5·4. *Note.* Even the rather weak condition of $\mathscr{T}^+$-compactness used in this section is not in general a necessary one. For example, the set $A = \{x\colon x_0 = 0\}$ is not $\mathscr{T}^+$-compact in the restriction of $\mathbf{L}^2$ to $\{x\colon |x|_1 < 1\}$; but any two causally ordered points belong to a chain which is future-connecting for A.

5·5. *Note.* We have no examples of causal spaces which are future-conjunct or -connected without being setwise so.

6. *Examples.* Throughout this section '$\mathbf{N}$' denotes the set of integers and '$\mathbf{R}$' the set of real numbers; $\mathbf{R}^{\pm} = \{r\colon \pm r > 0\}$; R^n is the Euclidean space of n-tuples

$$x = (x_0, x_1, \ldots, x_{n-1}).$$

We shall call a causal space *self-dual* if it is causally isomorphic (see 0·2) to its dual. (A self-dual space is future-connected if and only if it is past-connected, etc.)

The manifolds in Examples 6–1 to 6–7 are all equipped with their natural differentiable structure; and, when given the natural causal structure corresponding to the specified choice and time-labelling of null-cones, they are full and regular $\mathfrak{C}$-spaces.

Example 6–1. Let X be the cylinder obtained from the subspace $\{x: |x_1| \leqslant 1\}$ of $\mathbf{R}^2$ by identifying the two points $(x_0, -1)$ and $(x_0, 1)$ for each x_0; and define null-cones on X by the metric form $dx_0^2 - dx_1^2$. We time-label the null-cones continuously and give X its natural causal structure.

Although in this space any two beams intersect either infinitely often or not at all, it is completely non-pathological: it is a self-dual future-reflecting $\mathfrak{A}$- and $\mathfrak{B}$-space, and therefore both future-distinguishing and setwise future-connected.

This manifold is a 2-dimensional counterpart of the Einstein static universe.

Example 6–2. Let X be the punctured cylinder obtained from the subspace $\{x: |x_1| \leqslant 1\} \sim \{(0,0)\}$ of $\mathbf{R}^2$ by identifying the two points $(x_0, -1)$ and $(x_0, 1)$ for each x_0; and define null-cones on X by the metric form $dx_0 dx_1 - x_0^2 dx_1^2$. We time-label the null-cones by requiring the tangent vector $\partial/\partial x_0$ at each point of X to be future-directed, and give X its natural causal structure.

Writing $S^{\pm}$ and T for the intersection of X with, respectively, the sets $\mathbf{R}^{\pm} \times \mathbf{R}$ and $\{0\} \times \mathbf{R}$, each point of S^- (resp. T) chronologically precedes each point of T (resp. S^+), and T is a beam. The set $G = \mathbf{R} \times \{\frac{1}{2}\}$, though a null geodesic, is the union of the one-point set $G \cap T$ with the two disjoint beams $G \cap S^+$ and $G \cap S^-$: indeed T intersects no beam except itself.

This self-dual space is evidently not an $\mathfrak{A}$-space; it is future-reflecting, but, as it is not weakly distinguishing, it is not a $\mathfrak{B}$-space. It is nevertheless setwise future-connected. (To see this, note that, for example, the chain $X \cap (\mathbf{R} \times \{0\} \cup \{0\} \times \mathbf{R})$ is future-connecting and satisfies the conditions of Lemma 5–1.) On the other hand no chronological chain which intersects both T and S^+ is future-connecting.

If, using the horismos of this space, we construct an $\mathfrak{A}$-space on the same set X, it decomposes into three unrelated causal spaces on S^+, S^-, and T, of which each of the first two are causally isomorphic to subspaces of the space of Example 6–1, and the third to the space $(\mathbf{R}, \leqslant, \varnothing, \leqslant)$.

Example 6–3. Let X be the manifold obtained by deleting the points in $\mathbf{R}^- \times \{0\}$ from the manifold in Example 6–2, defining and labelling the null-cones at the remaining points as before; and let X have its natural causal structure.

Defining $S^{\pm}$, T, and G precisely as in Example 6–2 (with the new meaning for X), the null geodesic G is now the union of two beams: $G \cap S^+$ and $G \cap (S^- \cup T)$.

This is still not an $\mathfrak{A}$-space; but it is a past-reflecting $\mathfrak{B}$-space and therefore setwise past-connected. Although it is not future-distinguishing and therefore not future-reflecting, it is setwise future-connected. (To see this, note that, for example, the chain

$$X \cap (\mathbf{R}^- \times \{\tfrac{1}{2}\} \cup [T \sim \{(0,r): 0 < r < \tfrac{1}{2}\}] \cup \mathbf{R}^+ \times \{0\})$$

is future-connecting and satisfies the conditions of Lemma 5–1.) On the other hand no chronological chain which intersects both T and S^+ is future-connecting.

The $\mathfrak{A}$-space constructed from the horismos of this space on the same set X decomposes into a causal space on S^+, isomorphic to a subspace of the space of Example 6–1, together with a space on $S^- \cup T$, unrelated to the first and isomorphic to a subspace of $\mathbf{L}^2$.

Example 6–4. Let X be the Moebius band obtained from the subspace

$$\{x : 4|x_0| < 1,\ |x_1| \leqslant 1\}$$

of $\mathbf{R}^2$ by identifying the two points $(x_0, -1)$ and $(-x_0, 1)$ for each x_0. At all points for which $|x_1| < 1$ we define the null-cones by the metric form

$$(-dx_0^2 + dx_1^2)\cos\theta(x_1) + 2dx_0 dx_1 \sin\theta(x_1),$$

where θ denotes an odd differentiable monotone function on $\{r: -1 < r < 1\}$ into $\mathbf{R}$ such that $\theta(r) = \pi$ when $\frac{1}{2} < r < 1$. This metric has a continuous extension to the whole of X and defines null-cones which can be time-labelled continuously. The natural causal structure on X defines a self-dual, future-distinguishing $\mathfrak{A}$- and $\mathfrak{B}$-space which is not future-conjunct, and therefore not future-reflecting.

Example 6–5. Let X be the submanifold of $\mathbf{R}^2$ on $\mathbf{R}^2 \sim \{x: x_0 \geqslant 0,\ x_1 = 0\}$; define null-cones by the metric form $dx_0^2 - dx_1^2$, and time-label them by the rule used in Example 6–2. Equipped with its natural causal structure, X is an $\mathfrak{A}$- and $\mathfrak{B}$-space which is future- and past-distinguishing. It is future-reflecting and therefore setwise future-connected; but it is not past-conjunct and therefore not past-reflecting.

Example 6–6. Let X be the submanifold of $\mathbf{R}^3$ on $\mathbf{R}^3 \sim \{x: |x_0| \leqslant 1,\ x_1 = 0\}$, with null-cones defined by the metric form $dx_0^2 - dx_1^2 - dx_2^2$ and time-labelled continuously; and equip X with its natural causal structure.

This self-dual space is not future-reflecting or future-conjunct; it is an $\mathfrak{A}$-space, but not a $\mathfrak{B}$-space. However, since it is future-distinguishing, we can define a $\mathfrak{B}$-space on the same underlying set from its chronology. In the $\mathfrak{B}$-space the set

$$\{(r-1, r, p): r < 0\} \cup \{(r+1, r, q): r > 0\}$$

is a beam whenever $|p-q| \leqslant 2$; so the $\mathfrak{B}$-space is not regular, although it is still both an $\mathfrak{A}$- and $\mathfrak{C}$-space.

If the same method is applied to construct a $\mathfrak{B}$-space on the submanifold of X on

$$X \sim \{x: 2 \leqslant x_0 \leqslant 3, x_1 = 1\}$$

(using the chronology of this submanifold's natural causal structure), we obtain an $\mathfrak{A}$- but not a $\mathfrak{C}$-space.

Example 6–7. On $X = \mathbf{R}^2$ define causal relations by putting

$$x \ll y \quad \textit{if and only if} \text{ either } \quad x_0 < y_0 \quad \textit{and} \quad x_1 < y_1 \quad \text{or} \quad x_0 + 1 < y_0;$$

$$x \to y \quad \textit{if and only if} \text{ either } \quad x_0 = y_0 \quad \textit{and} \quad x_1 \leqslant y_1$$

$$\text{or} \quad x_0 \leqslant y_0 \leqslant x_0 + 1 \quad \textit{and} \quad x_1 = y_1.$$

Then X is a full and regular $\mathfrak{C}$- but not an $\mathfrak{A}$- or $\mathfrak{B}$-space. It is self-dual and future-reflecting, but not future-conjunct. Since it is also future-distinguishing, we can construct a $\mathfrak{B}$-space (which we shall call $X^{\mathfrak{B}}$) from its chronology. This is a regular $\mathfrak{A}$- and $\mathfrak{C}$-space which is future-conjunct and therefore (Theorem 5–4) setwise future-connected. (Note that the two beams through $(0, 0)$ are the sets $\{0\} \times \mathbf{R}$ and $\mathbf{R} \times \{0\}$ in X, but $\mathbf{N} \times \mathbf{R}$ and $\mathbf{R} \times \{0\}$ in $X^{\mathfrak{B}}$.)

(We might also seek to construct an $\mathfrak{A}$-, $\mathfrak{B}$-, and $\mathfrak{C}$-space on the same underlying set by another route. The $\mathfrak{A}$-space $(X, \prec^{\mathfrak{A}}, \ll^{\mathfrak{A}}, \rightarrow)$ determined by the original horismos is a $\mathfrak{B}$- but not a $\mathfrak{C}$-space. Forming the $\mathfrak{C}$-space whose causality is $\prec^{\mathfrak{A}}$ we do obtain an $\mathfrak{A}$- and $\mathfrak{B}$-space: but it is just an isomorph of $\mathbf{L}^2$.)

Since X, though Hausdorff in its Alexandrov topology, is not an $\mathfrak{A}$-space, its causal structure is not one naturally determined by any choice of null-cones on the manifold $\mathbf{R}^2$. (The same can be deduced for $X^{\mathfrak{B}}$ from the fact that it contains beams which are not connected in the manifold topology.) But a similar structure can be associated with a manifold 'naturally', provided we go to a higher number of dimensions.

In fact, X is a two-dimensional counterpart of the constant-amplitude anastigmatic plane-wave space-time (see (10)), which may be defined by the metric form

$$dx_0 dx_1 + \sum_{2}^{n-1} (x_i^2 dx_0^2 - dx_i^2)$$

on the manifold $\mathbf{R}^n$ $(n \geqslant 3)$. Equipped with its natural causal structure, this space shares all the causal properties we have listed for X, except that it is an $\mathfrak{A}$-space. Here $x \rightarrow y$ if and only if

$$\begin{aligned} &\text{either} \quad x_i = y_i \quad (i \neq 1) \quad \textit{and} \quad x_1 \leqslant y_1, \\ &\quad \text{or} \quad y_0 = x_0 + \pi \quad \textit{and} \quad x_i = y_i \quad (i \neq 0), \\ &\quad \text{or} \quad x_0 < y_0 < x_0 + \pi \quad \textit{and} \quad y_1 - x_1 = \Theta; \end{aligned}$$

and $x \ll y$ if and only if

$$\begin{aligned} &\text{either} \quad x_0 + \pi < y_0, \\ &\quad \text{or} \quad x_0 < y_0 \quad \textit{and} \quad y_1 - x_1 > \Theta; \end{aligned}$$

where

$$\Theta = \sum_{i=2}^{n-1} [(x_i^2 + y_i^2) \cos(y_0 - x_0) - 2x_i y_i] / \sin(y_0 - x_0).$$

The $\mathfrak{B}$-space constructed from this chronology has precisely the properties listed for $X^{\mathfrak{B}}$.

Example 6–8. Let p, q, a be points in $\mathbf{L}^n$ (where $n \geqslant 2$) satisfying $p \notin J^+(a)$, $q \in I^+(a)$, and $p \rightarrow q$. Then the line through p and q intersects the quadric hypersurface

$$C^+(a) \cup C^-(a)$$

in exactly one point, h; so that, if the coordinates of p; q; and a are all rational, so are those of h. Moreover, $h \in C^+(a) \cap [p, q]$; from which it follows that, if X is the restriction of $\mathbf{L}^n$ to the set of points all of whose coordinates are rational, then each link in X is future-connecting. Consequently X, which is clearly a self-dual, future-reflecting, full and regular $\mathfrak{A}$-, $\mathfrak{B}$- and $\mathfrak{C}$-space, is also, by Theorem 5–4, setwise future-connecting. Note, however, that no non-trivial chain in X is connected with respect to $\mathscr{T}^*$.

Example 6–9. Let X be the restriction of $\mathbf{L}^2$ to the subset

$$\{x: x \in \mathbf{N} \times \mathbf{N} \quad \textit{and} \quad x_0 + x_1 \quad \textit{is even}\}.$$

Then X is a regular $\mathfrak{A}$- and $\mathfrak{B}$-, but not $\mathfrak{C}$-space. It is not full; it is self-dual, future-reflecting, and setwise future-connected (Theorem 5–4). Using the causality of this

space to define a $\mathfrak{C}$-space on the same set, we obtain an $\mathfrak{A}$- and $\mathfrak{B}$-space, which is however, not regular.

Example 6–10. Let $\mathbf{S}$ denote an uncountable set and $\mathscr{F}$ the collection of its finite subsets. Let $\mathscr{X} = \mathscr{F} \cup \{\mathbf{S}\}$, and define causal relations on $\mathscr{X}$ by putting

$$P \prec Q \quad \textit{if and only if} \quad P \subset Q;$$
$$P \ll Q \quad \textit{if and only if} \quad P \in \mathscr{F} \quad \textit{and} \quad Q = \mathbf{S}.$$

It is easy to establish that $\mathscr{X}$ is future-conjunct for each of its subsets. On the other hand, if $\mathscr{R}$ is any chain in $\mathscr{F}$, the union of the members of $\mathscr{R}$ is countable and therefore a proper subset of $\mathbf{S}$. From this it follows that $\mathbf{S}$ itself can belong to no non-trivial future-connecting chain: that is, $\mathscr{X}$ is not future-connected.

We may, however, observe that $\mathscr{X}$, although future- and past-reflecting, is not weakly distinguishing, not full, not regular, and neither an $\mathfrak{A}$-, $\mathfrak{B}$-, or $\mathfrak{C}$-space.

REFERENCES

(1) Ahmavaara, Y. The structure of space and the formalism of relativistic quantum theory. I. *J. Math. Phys.* **6** (1965), 87–93.

(2) Bohm, D. A proposed topological formulation of the quantum theory. *The scientist speculates* [editor I. J. Good], pp. 302–314. (Heinemann; London, 1962).

(3) Coxeter, H. S. M. and Whitrow, G. J. World-structure and non-Euclidean honeycombs. *Proc. Roy. Soc. London Ser. A* **201** (1950), 417–437.

(4) Finkelstein, D. and Misner, C. W. Some new conservation laws. *Ann. Physics* **6** (1959), 230–243.

(5) Hawking, S. W. The occurrence of singularities in cosmology. *Proc. Roy. Soc. London Ser. A* **294** (1966), 511–521.

(6) Hill, E. L. Relativistic theory of discrete momentum space and discrete space-time. *Phys. Rev.* **100** (1955), 1780–1783.

(7) Misner, C. W. The flatter regions of Newman, Unti, and Tamburino's generalized Schwarzschild space. *J. Math. Phys.* **4** (1963), 924–937.

(8) Penrose, R. Gravitational collapse and space-time singularities. *Phys. Rev. Lett.* **14** (1965), 57–59.

(9) Penrose, R. Zero rest-mass fields including gravitation: asymptotic behaviour. *Proc. Roy. Soc. London Ser. A* **284** (1965), 159–203.

(10) Penrose, R. A remarkably property of plane waves in general relativity. *Rev. Mod. Phys.* **37** (1965), 215–220.

(11) Schild, A. Discrete space-time and integral Lorentz transformations. *Canad. J. Math.* **1** (1949), 29–47.

(12) Snyder, H. S. Quantized space-time. *Phys. Rev.* **71** (1947), 38–41.

(13) Zeeman, E. C. Causality implies the Lorentz group. *J. Math. Phys.* **5** (1964), 490–493.

(14) Bass, R. W. and Witten, L. Remark on cosmological models. *Rev. Mod. Phys.* **29** (1957), 452–453.

Document generated by Anna’s Archive around 2023-2024 as part of the DuXiu collection (https://annas-blog.org/duxiu-exclusive.html).

Images have been losslessly embedded. Information about the original file can be found in PDF attachments. Some stats (more in the PDF attachments):

{
  "filename": "NDA3ODg1NDEuemlw",
  "filename_decoded": "40788541.zip",
  "filesize": 126499015,
  "md5": "59684ce2bf301604f7f32693368b8b5f",
  "header_md5": "88ebb865e89acd170eb9065d5ec6bc47",
  "sha1": "91036fc994f29cd8b1f1f27cc1d138df2eea8d49",
  "sha256": "3c29078d2a3e481d49c33672fde37002e128cb9d33c99b39dd39c371007b8b28",
  "crc32": 1641033855,
  "zip_password": "",
  "uncompressed_size": 162277021,
  "pdg_dir_name": "ROGER PENROSE COLLECTED WORKS VOLUME 1 1953-1967_40788541",
  "pdg_main_pages_found": 822,
  "pdg_main_pages_max": 822,
  "total_pages": 845,
  "total_pixels": 5193606600,
  "pdf_generation_missing_pages": false
}